NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DETECTING AGE IN ONLINE CHAT

by

Jenny K. Tam

September 2009

Thesis Advisor:	Craig H. Martell
Second Reader:	Kevin M. Squire

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE September 2009	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE Detecting Age in Online Chat	5. FUNDING NUMBERS
6. AUTHOR(S) Jenny K. Tam	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

Over 90% of teens in the United States use the Internet, and many use it for social interaction. Due to the faceless nature of digital communication, criminals can easily pose as legitimate users to build friendship and trust with potential victims.

Even though fewer youths are going to chat rooms and talking to people they do not know, the number of youths receiving aggressive solicitations for offline contact has not declined. Most sexual solicitations go unreported to law enforcement and parents.

Though it is a crime for an adult to sexually exploit a minor, it is not always a crime for teens to solicit other teens. It would be of great help to law enforcement agencies if they could automatically detect adults soliciting teens versus teens soliciting other teens in online chat. This study analyzes the effectiveness of different machine learning techniques to distinguish chat conversation by teens and adults. Using proposed techniques, we classified teen and adult conversations with an accuracy of 86%. The goal of this research is to build an automatic recognition system of adults conversing with teens, capable of detecting predators and alerting agencies or parents of possible inappropriate conversations.

14. SUBJECT TERMS Authorship Profiling, Age Detection, Online Chat, Naïve Bayes Classifier, Support Vector Machine	15. NUMBER OF PAGES 147
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

DETECTING AGE IN ONLINE CHAT

Jenny K. Tam
Major, United States Army
B.S., United States Military Academy, 1998

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
September 2009

Author: Jenny K. Tam

Approved by: Craig H. Martell, PhD
Thesis Advisor

Kevin M. Squire, PhD
Second Reader

Peter J. Denning, PhD
Chairman, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Over 90% of teens in the United States use the Internet, and many use it for social interaction. Due to the faceless nature of digital communication, criminals can easily pose as legitimate users to build friendship and trust with potential victims.

Even though fewer youths are going to chat rooms and talking to people they do not know, the number of youths receiving aggressive solicitations for offline contact has not declined. Most sexual solicitations go unreported to law enforcement and parents.

Though it is a crime for an adult to sexually exploit a minor, it is not always a crime for teens to solicit other teens. It would be of great help to law enforcement agencies if they could automatically detect adults soliciting teens versus teens soliciting other teens in online chat. This study analyzes the effectiveness of different machine learning techniques to distinguish chat conversation by teens and adults. Using proposed techniques, we classified teen and adult conversations with an accuracy of 86%. The goal of this research is to build an automatic recognition system of adults conversing with teens, capable of detecting predators and alerting agencies or parents of possible inappropriate conversations.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I owe my success to the people who helped me complete this thesis. I would first like to thank my thesis advisor, Professor Craig Martell, for helping me find a topic of great interest, and giving me the freedom to explore different approaches. Your encouragement, guidance, and expertise were invaluable. I would additionally like to thank Professor Kevin Squire for your instruction and help with the more technical aspects of machine learning.

Special recognition goes to Ms. Jane Lin, my predecessor in this line of research, without whom my research would have been considerably more arduous. Your work to compile the NPS Chat Corpus and recommendations for future work were the springboard for my research.

To my parents, thanks for guidance and support given to me over the years.

Many thanks to my friends and lab partners, LT Jonathan Durham and Capt David Dreier. Jon, your suggestions, when I found myself at an impasse, were invaluable—especially the ones suggesting we all go watch a movie. Dave, your coding assistance was of tremendous help. You are an expert, and from you I learned a short answer could be long. After numerous group projects, I have come to realize, no power in the verse can stop either of you.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. TEENS AND THE INTERNET

In the United States, 93% of teens use the Internet mostly to gather information. Teens are also increasingly using the Internet for social interaction. In the Pew Internet & American Life Project Survey in 2006, 68% of the teens surveyed sent or received instant messages (IM) and 18% visited chat rooms. The decrease of chat room usage from 55% in 2000 may be due to increased awareness of the possible dangers, such as sex crimes, in chat rooms [1].

B. INTERNET SAFETY AND LAW ENFORCEMENT

The second Youth Internet Safety Survey (YISS-2), conducted in 2005 by Dr. Janis Wolak, Dr. Kimberly Mitchell, and Dr. David Finkelhor, funded by the National Center for Missing & Exploited Children, found that while there was a decrease in the proportion of youth receiving solicitations on the Internet, the number of dangerous sexual overtures or aggressive solicitations has not declined [2]. The study considered aggressive solicitations to be solicitations that involved offline contact via mail, telephone or in person or attempts/requests to meet offline. Education and law enforcement may have "deterred the casual solicitors, but the not more determined or compulsive solicitors [3]." Table 1 contains some of the statistics derived from the YISS-2 survey. It is interesting to note that in a majority of incidents, including aggressive incidents, the solicitors were younger than 18-years-old. Also, in 35% of the aggressive episodes, youths did not think the solicitations were serious enough to tell anyone. If they were to tell someone, they were more likely to tell a friend or sibling (29%), rather than a parent (18%). Only 7% of the aggressive solicitations were reported to law enforcement, an Internet Service Provider (ISP), or other authority.

Episode Characteristics	All Incidents (n=216)	Aggressive Incidents (n=68)
Age of Offender		
Younger than 18 Years	43%	44%
18 to 25 Years	30%	34%
Older than 25 Years	9%	15%
Don't know	18%	7%
Incident Known or Disclosed to		
Friend or Sibling	26%	29%
Parent/Guardian	12%	18%
Other Adult	2%	3%
Teacher, Counselor, or other School Personnel	2%	6%
Law Enforcement, ISP, or other Authority	5%	7%
Someone Else	4%	6%
No One	56%	35%
Of Youth Who Did not Tell Anyone, Why Didn't	56% (n=120)	35% (n=24)
Not Serious Enough	69%	71%
Afraid	13%	8%
Thought Might Get in Trouble	9%	8%
Other	6%	13%

Table 1. YISS-2 Internet Sexual Solicitation of Youth (n=1500) [After 2].

C. MOTIVATION

Given that a majority of youths do not think the aggressive solicitations are serious enough to report to an adult, let alone law enforcement, it is vital that juveniles are educated about the seriousness of the crime. If such education is not available, a method of notifying an adult, ISP, or law enforcement when a solicitation occurs would be of great help towards preventing a solicitation from turning into something more dangerous.

Offenders use a teenager's "developmentally-driven desire for romance and interest in sex to manipulate them into meeting them in person" [2]. Though most offenders do not deceive victims by posing as another youth, most youths

are not certain, or only somewhat aware, of the solicitor's age. Not all solicitors are adults. The YISS-2 survey found that there are a substantial number of peer solicitations [2].

To catch online predators, law enforcement officers or volunteers pose as youths in online chat rooms. Given the limited number of law enforcement officers and volunteers, an automated system that detects adults soliciting youths would augment the efforts of law enforcement officials. ISPs could also add such a system to parental control features [4].

Though it is a crime for adults to exploit a child sexually, depending on state laws, it is not always a crime for teens to solicit other teens. When analyzing suspicious chat behavior, it is important that law enforcement officials be able to separate the lesser number of cases of adults soliciting teens from the much greater number of cases of teens soliciting teens. The purpose of this thesis is to determine the effectiveness of different machine learning techniques in detecting teen chat posts from adult chat posts. The goal of this work is to facilitate an automatic recognition system of adults conversing with teens.

D. ORGANIZATION OF THESIS

This thesis is organized as follows:

- Chapter I discusses the role of the Internet and teens, Internet safety and the motivation for an automated system that can detect adults conversing with teens in online chat.
- Chapter II contains background information about chat, prior research in chat analysis and machine learning techniques used in this study.
- Chapter III explains in detail this study's approach to experiments, to include the source of data, classification tasks, feature selection and the setup of the experiments.
- Chapter IV contains the results of the experiments and analysis of the results.
- Chapter V contains concluding remarks and possible areas of future research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. ONLINE CHAT

The Internet Relay Chat (IRC) protocol allows large groups of people to converse with each other over the Internet. IRC uses a network of servers that relay messages to each other. Users connected to one server can communicate with other users on different servers in the same network. These networks can contain many chat rooms, known as "channels," and each chat room can contain thousands of users. Chat rooms can be public, where all users can read all chat messages, or "private," where users write directly to other specific users only. Some of the most popular chat programs are AOL Messenger, Yahoo Messenger, and MSN Messenger [5].

1. Chat Attributes

There are three components to a chat message. The chat participants are the first component. A screen name (made up or actual) identifies a participant, and is usually based on the participant's user profile. The second component contains optional information, such as a timestamp to identify when a user wrote a message. The last component is the chat message itself, usually displayed after the screen name of the user that typed that message [6]. See Figure 1 for an example of a chat dialog with component labels.

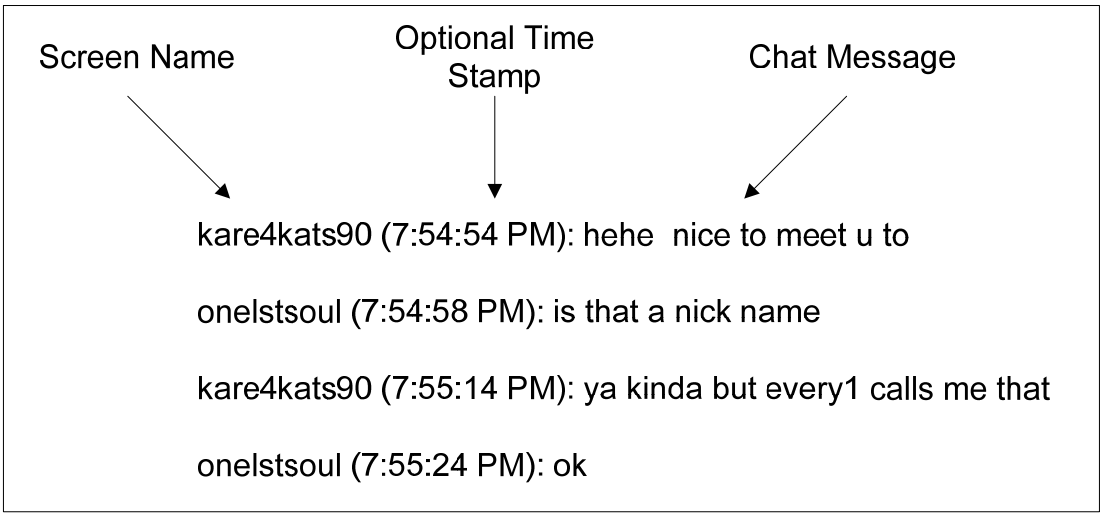

Figure 1. Components of a Chat Message.

2. Message Attributes

In [6], Haichao Dong, Aiu Cheung Hui, and Yulan He analyzed 72 pairs of MSN Messenger users from June to September 2005. Those users generated 33,121 messages during 1,700 conversation sessions. Dong, Hui, and He found that chat language is quite dissimilar from conventional English. Some of the features found in chat are acronyms, short forms, polysemes, synonyms, and misspelled words [6].

Acronyms are produced by taking the first letter in a sequence of words. *ASL* is the acronym for "Age Sex Location." Short forms are shortened forms of a word—*thx* for thanks. Polysemes are words that have more than one meaning. Synonyms are words that have the same or similar meaning and could replace each other. Misspelled words occur accidentally, but they could also be deliberate. Writers may write *noooooo* instead of *no* to give more emphasis to that word.

Additionally, messages contain icons, both textual and non-textual. Textual icons are known as emoticons, e.g., " :) " or " :(." Non-textual icons are graphics, such as a picture of a smiley face. Hyperlinks, mostly to other

websites, are also included in messages. Also, the length of messages tends to be short. In Dong, Hui and He's data set, 91.5% of chat messages are less than 50 bytes [6].

B. PRIOR WORK IN CHAT AUTHOR PROFILING

1. Author Profiling

Author profiling tries to determine an author's attributes, such as gender, age, educational background, and cultural background. Studies have shown that there are differences in communication by different ages, social groups, educational levels, and language backgrounds [7–9]. Therefore, it may be possible to model those differences, in order to detect the age of an author, based on the author's chat behavior.

2. Machine Learning and Text Analysis

"Machine learning is programming computers to optimize a performance criterion, using example data or past experience [10].” Past experience or sample data generates a pattern matching model that can make predictions about unseen data or future actions. The hope is that the model is a good approximation of the world, thus the predictions would be accurate. To measure how well a model performs, the predictions are measured against the actual truth in a controlled experiment. The evaluation measures are usually accuracy, precision, recall, and F-score, which are determined by using the number true positives, true negatives, false positives, and false negatives. Section C.4 contains a more detailed discussion of evaluation measures.

Machine learning lends itself well to text classification problems. There are many different machine learning algorithms and some of the more common ones used to classify text are Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and *k* Nearest Neighbor (*k*-NN) [11–13]. These techniques are described in [10, 14, 15].

Machine learning algorithms build classification models using statistical analysis of features in text. There are four categories of features—lexical, syntactic, structural, and content specific [13]. See Table 2 for examples of each.

Feature Category	Examples	
Lexical	Average word/sentence length	Vocabulary richness
Syntactic	Frequency of function words	Use of punctuation
Structural	Paragraph length	Use of indentation
	Use of a greeting statement	Use of a farewell statement
Content-specific	Frequency of key words	

Table 2. Categories of Features [From 13].

Within each category, there are many types of measures. In 1998, Joseph Rudman estimated authorship analysis applications have used nearly 1,000 different writing style features [16]. Many measures use word tokens and word types. Types represent the number of unique words in a corpus or vocabulary and tokens are the total number words in a corpus or vocabulary.

3. Prior Analysis of Chat Logs

Jane Lin tried to determine the gender and age group of an author using a NBC [5]. The dataset that she gathered is the same dataset used for this study. The dataset is a superset of the NPS Chat Corpus, a corpus of chat logs by authors of different ages. During her data preprocessing, she removed documents written by authors of an unknown age. She kept punctuation marks and did not perform stemming, which converts words to their stem (e.g., changing the words *running*, *runs*, *ran*, *runs* to *run*). Table 3 contains a summary of her training set data. Her corpus is described in more detail in Chapter III.A.1.

Corpus Breakdown	Number of Items
Author	2232
Tokens	658668
Types	72104
Sentences	143734
Sentences per Author	64.4
Tokens per Sentence	4.6

Table 3. Lin's Training Set [After 5].

She used the following features:

- Emoticon token counts
- Emoticon types per sentence
- Punctuation token counts
- Punctuation types per sentence
- Average sentence length
- Average word type count per document

Appendix F contains her emoticon and punctuation dictionary.

She set aside 10% of her data as the test set. Table 4 shows a breakdown of the corpus by age group.

Category	Training Set	Test Set
Teens (13-19)	591	68
20s (20-29)	882	97
30s (30-39)	355	37
40s (40-49)	301	32
50s (50-59)	103	10
Adult (20-59)	1641	176

Table 4. Division of Data in Lin's Experiments by Age Group [After 5].

Her experiments to classify teens versus 20-year-olds failed to generate notable results. As she compared teens against older and older age groups, however, her results monotonically increased until generating an F-score measure of 0.932 for teens against 50-year-olds [5]. In her calculations, she

used a NBC that included a prior probability and a NBC that did not use the prior probability. Tables 5 and 6 contain the best F-score results for each of her classification tasks.

Classification Task	Precision	Recall	F-score
Teens vs. 20s	0.857	0.088	0.160
Teens vs. 30s	0.648	1.000	0.786
Teens vs. 40s	0.687	1.000	0.814
Teens vs. 50s	0.872	1.000	0.932
Under 26 vs. 26 and Older	0.541	1.000	0.702

Table 5. Lin's Results from using a NBC with Prior Probability [After 5].

Classification Task	Precision	Recall	F-score
Teens vs. 20s	0.422	0.515	0.464
Teens vs. 30s	0.663	0.926	0.773
Teens vs. 40s	0.684	0.956	0.798
Teens vs. 50s	0.865	0.941	0.901
Under 26 vs. 26 and Older	0.530	0.947	0.679

Table 6. Lin's Results from Using a NBC without Prior Probability [After 5].

Lin concluded that her results were influenced by the prior probability. The NBC would predict a test case to be a member of the class with the highest prior probability, which is based on the proportion of the class in the training data. Though she ran experiments without a prior, the predictions had only slightly better F-scores or precision. Given her results and the characteristics of the data, she believed that other machine learning techniques, such as a SVM, or bigrams or higher order n-grams might generate better results [5].

Kucukyilmaz et al. used different machine learning techniques to predict user and message attributes based on online chat and their work is described in [11]. They compared the following machine learning algorithms: *k*-NN, NBC, SVM and Patient Rule Induction Method (PRIM). The attributes predicted were gender, age, school, connection domain, receiver, author identity, and day

period. The chat corpus they used was from an inactive chat server, Heaven BBS, and the messages were in the Turkish language. They used both term features, and stylistic features. Their term-based features were single tokens, unigrams from the vocabulary of the corpus. Table 7 contains the stylistic features used and possible feature values.

Feature	Features in the Category	Possible Feature Values
Character usage	Frequency of each Character	Low, Medium, High
Message length	Average Message Length	Short, Average, Long
Word length	Average Word Length	Short, Average, Long
Punctuation Usage	Frequency of Punctuation Marks	Low, Medium, High
Punctuation Marks	A List of 37 Punctuation Marks	Exists, Not Exists
Stopword Usage	Frequency of Stopwords	Low, Medium, High
Stopwords	A List of 78 Stopwords	Exists, Not Exists
Smiley Uages	Frequency of Smileys	Low, Medium, High
Smileys	A List of 79 Smileys	Exists, Not Exists
Vocabulary Richness	Number of Distinct Words	Poor, Average, Rich

Table 7. Features Used in Chat Research by Kucukyilmaz et al. [From 11].

In their term-based experiments, they performed three preprocessing steps on the data. The first step cleaned and filtered the data. They removed single word messages, non-alphanumeric characters and terms that were on a list of 78 Turkish stop words. They calculated the term frequency-inverse document frequency (tf-idf) values for the remaining terms and used those values, also known as tf-idf weighting[1], as the feature values. If the users had less than a pre-determined number of terms, those authors were removed.

[1] $tf-idf\ \ weighting = tf_{i,j} \times idf_i$. $tf_{i,j} = \frac{w_{i,j}}{\sum_k w_{k,j}}$, where $w_{i,j}$ is the number of times term i appears in document j and $w_{k,j}$ is the total number of terms in document j.
$idf_i = \log(N \div n_i)$, where N is the total number of documents in the corpus and n_i is the number of documents in which term i appears. See [14] for more information.

In the second step, Kucukyilmaz et al. balanced the data sets by selecting from each class an equal number of instances with the highest term counts. To balance the instances, each instance was limited to 3,000 terms, and they discarded any additional terms.

In the third step, to reduce the dimensionality of the data set, they retained the most discriminative features and discarded the less discriminative features. They did this by apply the χ^2 (CHI square) statistic[2] to every term to calculate each term's discriminative power.

In their style-based experiments, they did not remove punctuation marks or stop words, which are words that have syntactic functions, but do not contribute to content. Given each user had an almost equal number of features, they did not perform any instance balancing.

Table 8 contains the breakdown of their corpus. Kucukyilmaz et al. did not state the number of tokens in their corpus. Based on the number of authors reported, chat posts and average number of words per post, they had approximately 1,603,072 tokens.

Corpus Breakdown	Number of Items
Author	1616
Tokens	1603072 (approximate)
Types	165137
Posts	218742
Posts per Author	160
Words per Post	6.2

Table 8. Kucukyimaz, et al. Corpus Breakdown [11].

For their experiments predicting birth year before 1976 (inclusive) and after 1976 (exclusive), they had 60 authors (30 prior to 1976 and 30 after 1976) in their test set. At the time of their study, people born after 1976, were older

[2] $\chi^2 = \sum_n \frac{(Obs - Exp)^2}{Exp}$, where Obs are observed frequencies of a term and Exp are the expected frequencies of a term. See [15, 27] for more information.

than 24-years old. In their experiments predicting birth year (1975, 1976, 1977, 1978), they had 30 authors per year in their test set (120 authors total). They performed 10-fold cross validation, where they set aside a different 10% (with replacement) of their corpus data for their test set; they ran all experiments 10 times, using a different training/test set each time. Their measure of performance for each classifier was accuracy. Tables 9 and 10 show the average accuracy results for their term-based and style-based classification experiments.

Classification Task	*k*-NN	Naïve Bayes	PRIM	SVM
Birth Year <= 1976 (<= 24 years old)	50.1	60.8	53.8	56.3
Birth Year (1975, 1976, 1977, 1978)	24.0	27.3	20.0	26.5

Table 9. Term-Based Classification Accuracy Results by Kucukyilmaz et al. [After 11].

	k-NN	Naïve Bayes	PRIM	SVM
Birth Year <= 1976 (<= 24 years old)	50.0	75.4	55.5	48.0
Birth Year (1975, 1976, 1977, 1978)	22.8	37.4	19.9	22.0

Table 10. Style-Based Classification Accuracy Results by Kucukyilmaz et al. [After 11].

They concluded that word choice and writing behavior could predict characteristics of chat users and messages. Like chat by English language users, chat by Turkish language users had frequent slang words and misspellings. Style-based feature sets were more effective than term-based features in determining the birth year of an author. In their data set, younger users more often had a smaller vocabulary and they preferred using emoticons more than older users [11].

4. Analysis of Perverted Justice Chat Logs

Nick Pender used a SVM model and a *k*-NN classifier to detect online sexual predators [4]. His corpus consisted of online chat conversations between

sexual predators and pseudo victims, who were adult volunteers posing as underage victims. These volunteers made their conversations with subsequently convicted predators available on the Perverted Justice website[3].

His corpus contained 701 text logs, where each log contained all the conversations between one pseudo victim and the predator pursuing that victim. The log sizes were between 269 and 42,220 words, including timestamps and screen names. The corpus contained 2,603,681 words, including screen names, timestamps, misspelled words, and punctuation marks. Pender divided each text log into two files: one file containing all the victim chat posts and another file containing all the non-victim chat posts. After dividing the text logs, he had a corpus of 1,402 files. His training set contained 1,122 files (561 victim/561 non-victim) and the test set contained 280 files (140 victim/140 non-victim) [4].

He used a bag of unigrams, bigrams, and trigrams separately as features. In a bag of n-grams, token order is irrelevant, so the trigram *the cat sat* is the same trigram as *sat the cat*. Data preprocessing included removal of stop words based on the 79 most frequent word types in his corpus. He further reduced dimensionality by performing feature reduction using a combination of document frequency and odds ratio of terms. He created nine feature sets by selecting 5,000, 7,500, and 10,000 unigrams, bigrams, and trigrams with the highest average odds ratios. For each document in the training and test set, he calculated the tf-idf values of the extracted features. Thus, each document had nine different representations, depending on whether unigrams, bigrams, or trigrams were the features and the dimensionality of each vector (5,000, 7,500, 10,000) [4].

The *k*-NN classifier used k values of 5, 10, 15, 20, 25, or 30. The SVM model used a linear kernel. See Table 11 for results of the *k*-NN classifier and Table 12 for results of the SVM model.

[3] Freely available from Perverted Justice at http://www.perverted-justice.com/guide/.

Terms	*k* Value	F-score		
		5000	7500	10000
Unigram	5	0.546	0.586	0.814
	10	0.675	0.571	0.854
	15	0.607	0.575	0.818
	20	0.586	0.561	0.811
	25	0.579	0.582	0.818
	30	0.571	0.575	0.807
Bigram	5	0.500	0.500	0.500
	10	0.500	0.511	0.500
	15	0.582	0.500	0.500
	20	0.514	0.500	0.500
	25	0.504	0.500	0.675
	30	0.500	0.500	0.779
Trigram	5	0.504	0.500	0.514
	10	0.504	0.500	0.871
	15	0.500	0.532	0.925
	20	0.504	0.507	0.918
	25	0.529	0.500	0.936
	30	0.511	0.500	0.943

Table 11. F-score of the *k*-NN Classifier with Different *k* Values and Dimensions [From 4].

Feature	F-score		
	5000	7500	10000
Unigram	0.558	0.415	0.415
Bigram	0.575	0.545	0.545
Trigram	0.893	0.908	0.908

Table 12. F-score of the SVM Models with Different Dimensions [From 4].

Pender found that with both classifiers, trigrams with the highest dimension vector performed the best. The *k*-NN classifier performed slightly better than the SVM. When using bigrams, only *k* values of 25 and 30 and using 10,000 bigrams produced results better than chance. This led Pender to suggest that words may not be enough to distinguish the conversations, but that word phrases (i.e., how words are put together) distinguish conversations [4].

C. MACHINE LEARNING TECHNIQUES

Though there are many different machine learning algorithms, this paper focuses on the two algorithms used in this research, the Naïve Bayes Classifier and the Support Vector Machine.

1. Naïve Bayes Classifier

The Naïve Bayes Classifier (NBC) is discussed in [5, 10, 14, 15]. It uses Bayes' Theorem and makes strong independence assumptions among the data being classified. While these assumptions are almost always false, because of them, this probabilistic classifier is very easy to implement.

Bayes' Theorem is used to derive the conditional probability of an event, X, given Y, based on the probabilities of X and Y and the probability of Y, given X. Bayes' Theorem is written as

$$P(X \mid Y) = \frac{P(X)P(Y \mid X)}{P(Y)}.$$

The NBC assumes *n* random variables for a feature vector, F_n, of $f_1, \ldots, f_n$ features and a random variable, C, for classes, the probability of which is conditional on the set of features. Combined with Bayes' theorem, we get

$$P(C \mid \mathrm{F}_n) = \frac{P(C)P(\mathrm{F}_n \mid C)}{P(\mathrm{F}_n)}.$$

If the set of potential classes (i.e., teen, 20s, 30s, 40s, 50s, or adult), C, is known, then most probable class, c^*, given F_n, is the one with the highest probability, or

$$c^* = \arg\max_{c_i \in C} \left[\frac{P(\mathrm{F}_n \mid c_i)P(c_i)}{P(\mathrm{F}_n)} \right].$$

Because the term $P(\mathrm{F}_n)$ does not change between classes, the *argmax* operator allows one to discard it. The formula then becomes

$$c^* = \arg\max_{c_i \in C} \left[P(\mathrm{F}_n \mid c_i) P(c_i) \right].$$

The NBC assumes independence among the features, meaning that each element in the feature vector, F_n, is independent to every other element in the vector. This means that

$$P(\mathrm{F}_n \mid c_i) = \prod_{f_j \in \mathrm{F}_n} P(f_j \mid c_i)$$

so,

$$c^* = \arg\max_{c_i \in C} \left[P(c_i) \prod_{f_j \in \mathrm{F}_n} P(f_j \mid c_i) \right].$$

In this research, the features are vocabulary terms (e.g., an n-gram). Because the vocabulary size can be very large, the probability of a term appearing in the vocabulary can be quite small. Rather than using the probability of a feature, given an n-gram, the sum of the log of the probability is used to prevent numeric underflow so,

$$c^* = \arg\max_{c_i \in C} \left[\log P(c_i) + \sum_{f_j \in \mathrm{F}_n} \log P(f_j \mid c_i) \right].$$

2. Smoothing

One of the weaknesses of an NBC is its penalization of terms that do not appear in the training set, but do appear in the test set. Such terms, known as zero counts, will have a zero probability. To deal with that problem, smoothing is done to reapportion some probability mass from the more frequent terms to the

zero count terms [14]. There are many different smoothing algorithms and this research used two of them—Laplace and Witten-Bell smoothing.

a. Laplace Smoothing

By giving all token frequency counts an additional count, Laplace smoothing reapportions probability mass to the unseen terms. In an unsmoothed estimate, the probability of a term, t_i, is its count, c_j, normalized by the total number of tokens, N, in the vocabulary:

$$P(t_i) = \frac{c_i}{N}.$$

Since Laplace smoothing adds one to each term's count, the normalizing constant must be adjusted. Because each term's count has been increased by one, the normalizing constant's adjustment is the addition of V, the size of the vocabulary [14]. The smoothed estimate for a term would then be:

$$P_{Laplace}(t_i) = \frac{c_i + 1}{N + V}.$$

b. Witten-Bell Smoothing

Witten-Bell smoothing is discussed in [15, 17]. With this smoothing technique, a zero count event's probability is an estimate of the probability of seeing a new unseen event as one goes through the training set [15].

The probability for a non-zero count term, t_i, is found by the following formula, where c_i is the number of times t_i has appeared so far; n is the number of tokens seen so far; and v is the number of types that have appeared so far [17]:

$$P_{Witten-Bell}(t_i) = \frac{c_i}{n + v}.$$

The total probability for new unseen events is [17]:

$$P_{Witten-Bell}(t_{novel}) = \frac{v}{n+v}.$$

3. Support Vector Machine

The Support Vector Machine is discussed in [10, 18–21]. Also known as a maximum margin classifier, the SVM tries to find the line between two classes of data that maximizes the margin between them. Because data being classified is not always linearly separable (i.e., there does not always exist a line or hyperplane which can separate the two classes of data), the data is often transformed using a kernel function. Though there are different types of kernels for an SVM, this research used a linear kernel, which does not transform the data. The two classes of data (e.g., teens and adults), are represented by n-dimensional vectors, where each dimension represents a feature, such as an n-gram. Using the training set vectors, the SVM generates the hyperplane (model vector) that separates the two classes with the maximum margin. The test set vectors and model vector are then used to determine which side of the hyperplane the test vectors lay. The side that a test vector lies upon is its predicted class.

In Figure 2, both lines H_1 and H_2 separate the two classes, but line H_2 separates the classes with the maximum margin.

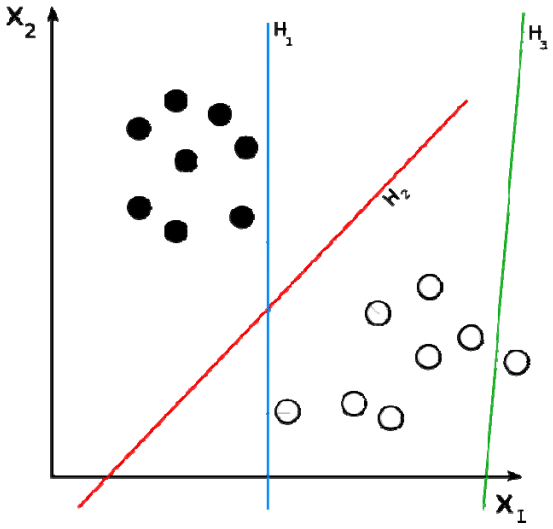

Figure 2. Linear Classification [From 22].

Based on the training data, a SVM will find the maximum margin hyperplane that separates the two classes. A maximum margin hyperplane exists where the distance from each class' closest data point to the hyperplane is as large as possible. Support vectors are the data points that are on the margin. Figure 3 is an example of a hyperplane that creates the maximum margin between classes. Also, in Figure 3, the support vectors are circled. The maximum margin in the figures is the distance between lines l_1 and l_2.

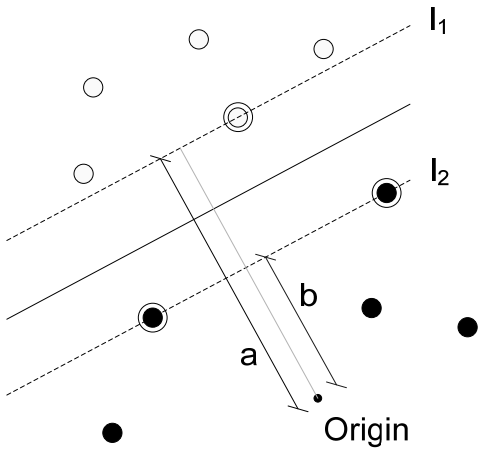

Figure 3. Linear Separating Hyperplanes.

Conversations between different groups can be very similar (e.g., between teens and 20s), thus classes are likely to overlap or have a very small margin. "Slack variables" compensate for this effect [21]. Appendix A provides further details on deriving the maximum margin hyperplane and the slack variables.

4. Measures of Classification Performance

a. Precision, Recall, and F-score

We used precision, recall and F-score measurements to evaluate the results of the experiments. Precision measures the correctness of the classifier by measuring the proportion of items the classifier correctly selected [15]. In other words, the percentage of chat posts the classifier predicted as authored by teens that were actually written by teens. A precision of 1.00 would mean that no adult chat posts were labeled as a teen chat post. Recall measures the proportion of all the targeted items the classifier selected [15]. In other words, did the classifier find all the teen chat posts? A recall of 1.00 would mean that all teen chat posts were found. The formulas for precision and recall are as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$TP =$ True Positive, number of posts correctly identified as authored by a teen

$FP =$ False Positive, number of posts incorrectly identified as authored by a teen

$FN =$ False Negative, number of posts incorrectly identified as authored by an adult

An F-score measure is used so that one cannot make trade-offs to favor precision at the expense of recall or vice versa. The F-score is the harmonic mean of precision and recall and the formula is as follows, where P is for precision and R is for recall:

$$F - score = \frac{2}{\frac{1}{P} + \frac{1}{R}}.$$

b. Accuracy

Accuracy is an often-used evaluation measurement. It calculates the percentage of items labeled correctly and the formula for accuracy is

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$TP =$ True Positive, number of posts correctly identified as authored by a teen

$FP =$ False Positive, number of posts in correctly identified as authored by a teen

$TN =$ True Negative, number of posts correctly identified as authored by an adult

$FN =$ False Negative, number of posts incorrectly identified as authored by an adult

Accuracy, however, may not be the best measure to use when the number of true negatives can be much greater than the number of true positives. Precision and recall are more sensitive to counts of true positives, false positives and false negatives, whereas accuracy is not [15]. If teens only wrote 10% of the documents in a 100-document corpus and the classifier correctly identified all of the documents not written by teens, but only one of the documents written by a teen, the accuracy would be 0.910. The F-score, however, would be 0.182. The accuracy measure is very good, but the classifier had missed 90% of the

documents of interest. In the above situation, if the classifier had not found any of the documents written by teens, the accuracy measure would be 0.900, but the F-score would be 0.000.

THIS PAGE INTENTIONALLY LEFT BLANK

III. TECHNICAL APPROACH

A. SOURCE OF DATA

1. Lin Chat Corpus

The chat data used was gathered in 2006 by Lin from a publicly available chat host and is described in [5]. Though the chat room server hosted scheduled chat rooms and chat rooms for different topics, Lin gathered data from chat rooms organized by age to keep topics as general and unbiased as possible. A portion of this data is available as the NPS Chat Corpus.[4]

In the Lin corpus, each chat log contains all the chat posts of a unique author; each log is labeled by the age of the author (self-reported in the author's profile information). The chat logs contain only the messages written by the author and do not contain time stamps or the author's screen name preceding his message. The corpus contains 3,290 unique authors.

2. Division of Data

We considered each chat log as a document, and each line in the chat log to be an individual post. All documents by authors of unknown age, and files with less than three words, were removed. Such short chat logs usually only contained greetings, and thus did not contain useful information. This left a total of 2,161 documents, each by a unique author, containing 292,831 posts comprised of 732,964 tokens and 85,479 word types. The test set contained 432 randomly selected documents (20% of total number of documents). This test set was not used for feature selection or training. See Table 13 for distribution of documents by age group.

[4] Available for non-commercial, non-profit educational and research use from The NPS Chat Corpus at http://faculty.nps.edu/cmartell/NPSChat.htm.

Category	Training Set	Test Set
teens (13-19)	465	116
20s (20-29)	689	172
30s (30-39)	259	65
40s (40-49)	235	59
50s (50-59)	80	20
Adult(20-59)	1263	316

Table 13. Number of Documents in the Training and Test Set.

B. CLASSIFICATION TASKS

Though the exact age of an author is known for each document, authors were placed into age groups. Our task was to determine which age group an author belonged. We performed a binary classification task between teens and a specific age group. There were five classification tasks:

- Teens versus 20-year-olds
- Teens versus 30-year-olds
- Teens versus 40-year-olds
- Teens versus 50-year-olds
- Teens versus adults (20–59-year-olds)

C. FEATURE SELECTION

Based on Pender's success with an SVM to classify pseudo teens and actual adults and Lin's recommendation, we choose not only to use her feature set using a SVM, but also to use higher order n-grams on both a SVM and NBC [5]. Kucukyilmaz et al. had some success using unigrams with a NBC and SVM, so our goal was to improve upon their results using higher order n-grams as well. In their stylistic feature set, they did not use actual counts of a feature's frequency, but rather threshold values, such as, poor/low/short, medium/average, rich/high/long, exists and not exists [11]. The experiments in this research explored if it would be more effective to use finer grain analysis; thus, meta-data feature values were the actual counts of appearances, rather than threshold values.

1. Features

The following word-based features were used: unigrams, bigrams, trigrams, and *character* trigrams. In addition, we used *character* 4-grams and 5-grams in the NBC experiments. To see if we could improve upon her results with a different classifier, we used Lin's feature set (Chapter II.3) in the SVM experiments. We also created a slightly different meta-data feature set (described below) for the SVM experiments.

Unigrams are single tokens (e.g., <*a*>, <*single*>, <*word*>). Bigrams contain two tokens, where order matters (e.g., <*two word*>, <*tokens are*>, <*bigram examples*>). Trigrams contain three tokens, where order matters (e.g., <*trigrams contain three*>, <*contain three words*>, <*words three contain*>).

We used *character* n-grams, a series of *n* characters, because they can capture indications of style including lexical information, contextual information, and use of punctuation. Additionally, such n-grams are noise tolerant. When texts contain grammatical errors or non-standard use of punctuation (e.g., emoticons), character n-grams are not as affected [12]. For example, the words *misspelled* and *mispelled* would generate many common character trigrams, but in a lexically-based representation, they would just be two different types. The character n-gram also captures errors that could be considered an identifying feature for a class (e.g., *ssp* and *spe*).

We measured the following meta-data for each document:

- Capital Letters–Average number of capital of letters per post. Measured by adding the total number of capitals in a document and dividing by the total number of posts in the document.
- Unigram Tokens–Average number of tokens per post. Measured by the total number of tokens and dividing by the total number of posts in the document. The posts were not stripped of punctuation or emoticons.

- Emoticon Types–Average number of emoticon types per post, as a measure of emoticons in a document. This was measured by the total number of emoticon types per post and dividing by the total number of posts in the document.
- Word Tokens–The average post length was measured by the total number of word tokens divided by the total number of posts. These word tokens were stripped of punctuation and emoticons.
- Word Types–Measure of richness of vocabulary. This was measured by adding all the word types in a document and then dividing by the total number of posts in the document.

In chat, it is not unusual for people to add letters to words to accentuate them, such as spelling the word *cool* as *cool* or *coooool*. Internet slang may misspell words as well. Instead of *cool*, some people spell that word as *kewl*. We felt that correcting spelling would remove features that would distinguish between different age groups. We also did not perform stemming, the process of reducing words to their stem (e.g., changing the words *running*, *runs*, *ran*, *runs* to *run*), for the same reason.

Both Lin and Kucukyilmaz et al. found that the younger the person, the more a post contained emoticons and emoticon types [5, 11]. Thus, we kept punctuation to maintain emoticons, but conducted all NBC experiments both with and without punctuation. The punctuation marks used are all the possible punctuation characters on a standard QWERTY keyboard. Appendix H contains the list of punctuation marks used in this research. All posts were converted to lower case letters to reduce the size of the dictionary. To account for the use of capital letters, we added that feature to the meta-data feature set described above.

The emoticons used in this study are from Wikipedia[5]. On that Web site, there are two emoticon styles—Eastern and Western. Eastern style emoticons, which originated in East Asia, can be read by a person without having to tilt his head [23]. An example of a smile is "(^_^)." Readers read Western style

[5] Freely available from Wikipedia at http://en.wikipedia.org/wiki/List_of_emoticons.

emoticons from left to right and usually tilt their head to read them. An example of a smile is ":-)." We limited our emoticons to the Western style type.

The Lin corpus also includes built-in emoticons specific to the chat server. Though there are more built-in emoticons that display colored icons, the list from Wikipedia contains the text representation of such emoticons. Table 14 contains the built-in emoticons we added to the dictionary of emoticons. Appendix G contains the full dictionary of emoticons used.

Emoticon
:beer:
:blush:
:love:
:tongue:

Table 14. Built-in Emoticons.

2. Stop Words

Typically, *stop words*, which have syntactic functions in English, but do not contribute to content, are removed from the vocabulary. We generated our own stop word list, because online chat communication has its own vocabulary and does not follow conventional spelling rules. We used two different methods of generating the *stop words* (actually *stop n-grams*). In both the mutual high-frequency and entropy-based methods, we removed *n* number of n-grams. For values of *n*, we used 5, 15, 25, 50, and 75. The n-grams were found by separating words using white space. Punctuation marks were retained and were kept attached to a word if there was not white space separating that word and the punctuation mark(s). Thus, the same word that was found in the middle of a sentence and at the end of a sentence was considered two different word types.

a. Mutual High-Frequency Stop Words

The first method, mutual high-frequency stop words, found the *n* most frequently used mutual n-grams between two classes. So if *n*=5, then the list would contain the five most popular mutual words in the teen and adult dictionaries.

To generate the stop n-gram list for each classification task, we first created age group specific n-gram dictionaries with the frequency count for each n-gram. All the dictionaries were then sorted by the frequency of use (token count) in descending order. For each classification task (e.g., teens versus 20s), using the sorted dictionaries for each age group, we found the *n* most mutually popular n-grams. To prevent n-grams from being unfairly chosen (e.g., a mutual n-gram has a very high frequency count in one class but a very low frequency count in the other class), n-grams selected were within the top 500 most frequent n-grams per class.

As an example, Figure 4 depicts the process to generate the five mutual high-frequency unigram list for the teens versus 50s classification task. The teen and 50-year-old unigram dictionaries are sorted by token count and the most frequent mutual tokens are selected for the stop n-gram list.

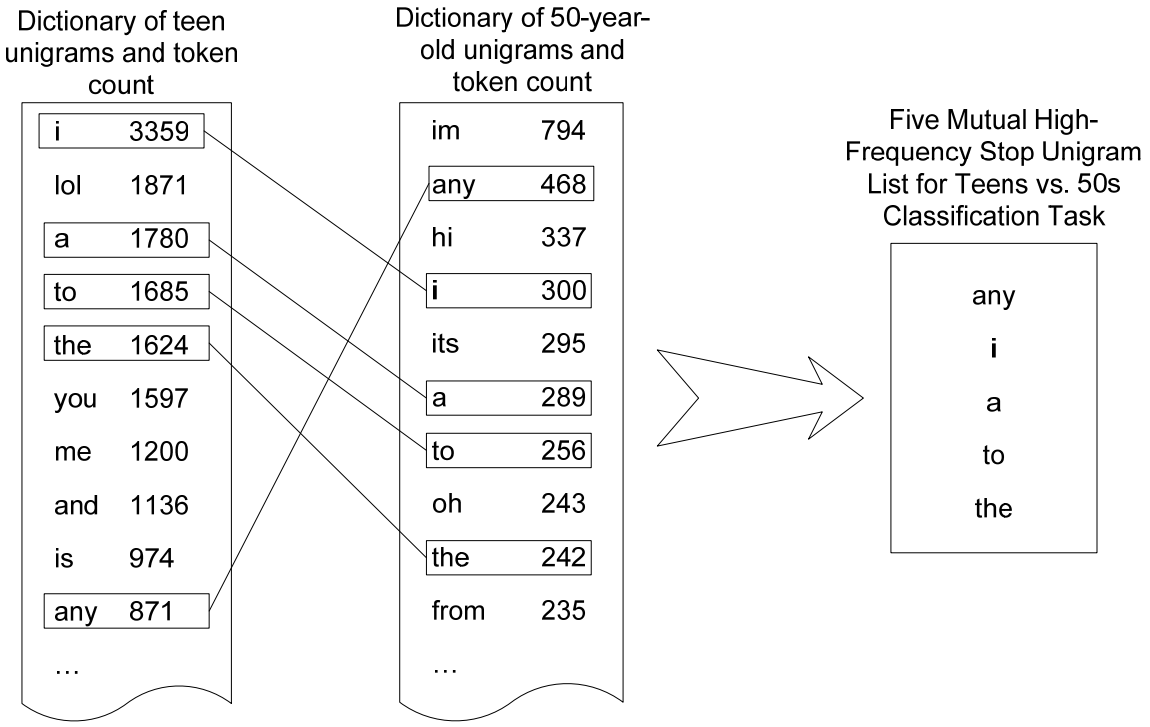

Figure 4. Creating a Mutual High-Frequency Stop N-gram List.

b. Entropy-Based Stop Words

The second method used entropy as a measure of information gain, that is, how much a given feature contributes to separating the training examples into their target classifications [24]. We used the following formula to measure the conditional entropy of $P(C \mid n_i)$, the probability of a class given n-gram i:

$$H(P(C \mid n_i)) = -\sum_j p(c_j \mid n_i) \log_2 p(c_j \mid n_i)$$

The higher the conditional entropy, the more equally distributed the n-gram is across the classes and therefore the less discriminative the n-gram is.

To generate the list for each classification task, we first created age group specific n-gram dictionaries with the frequency count for each n-gram. Next, for each classification task, we measured the conditional entropy using the

applicable age group dictionaries. The resultant n-grams were sorted in descending order by the entropy value. We then took the *n* number of n-grams needed from the rank ordered list.

We used this second method because we felt that mutual high-frequency-based stop n-grams might contain contextual information. Also, different age groups may use mutual n-grams, but one age group might use them more frequently than another. We wanted to see if there was a difference in performance using an n-gram's discriminative power versus its high frequency count.

D. EXPERIMENT SETUP

Figure 5 provides a summary of the process used to set up the experiments for each classification task.

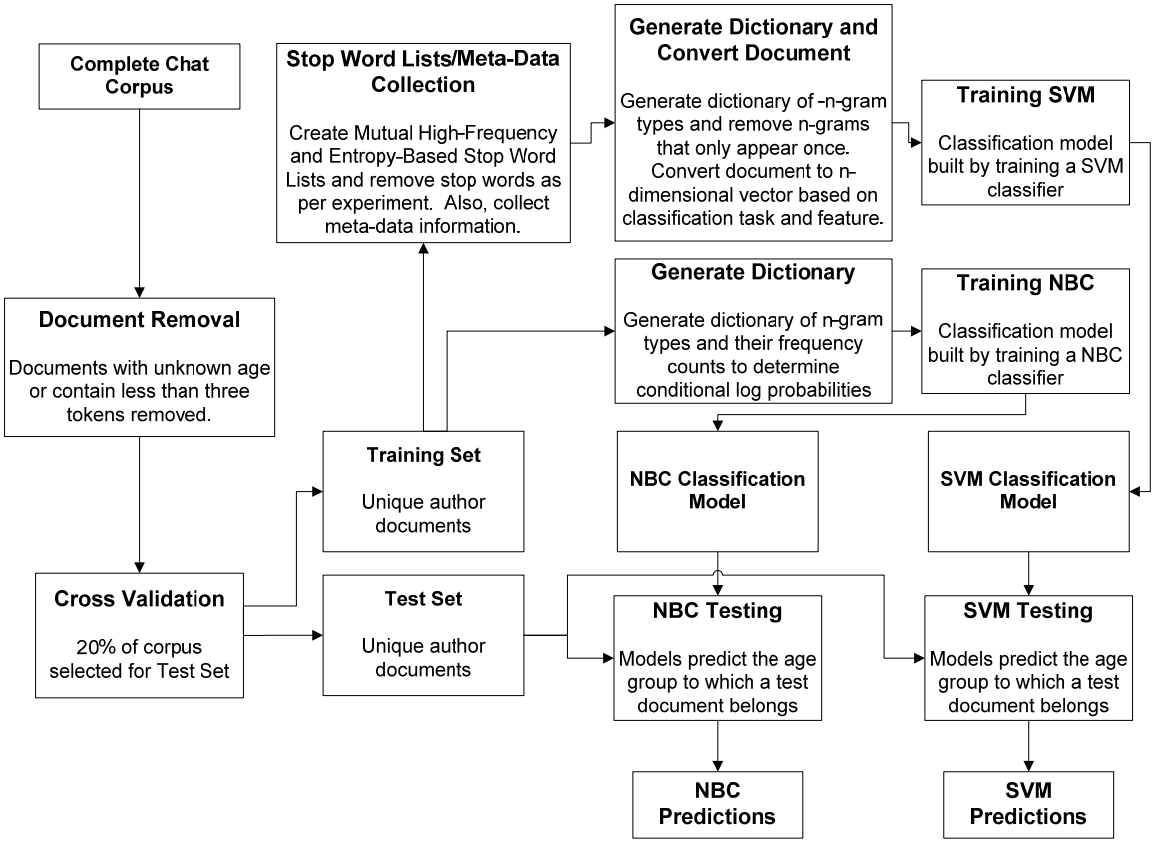

Figure 5. Experiment Process for each Classification Task.

1. Data Preprocessing

We made a pass over the data set for each n-gram size to generate the n-gram dictionary using a sliding window of one, two, and three tokens. Tokens were delineated by white space, except when generating character grams. Character grams included white space as a character token. When generating n-grams of size greater than one, we added a beginning of post and end of post tag; this captures information about the placement of an n-gram (e.g., how likely an n-gram starts or ends a post). Each n-gram dictionary only contained n-grams of n size (e.g., trigram dictionary only contained trigrams). Also, during each pass, meta-data features were calculated and stop n-grams removed. In the SVM training data, *hapax legomena*, n-grams that only appeared once were removed in each classification task to reduce dimensionality. In all the SVM experiments, punctuation marks were retained and were kept attached to a word if there was not white space separating that word and punctuation mark(s).

Table 15 shows the resultant number of unigram, bigram, trigram, and *character* trigram tokens (no stop words or entropy-based words removed, *hapax legomena* removed).

Category	Unigrams	Bigrams	Trigrams	3 Character Grams
Teens/Adults	575453	516741	212587	3160572
Teens/20s	323135	287568	115610	1768926
Teens/30s	213909	179526	68153	1187163
Teens/40s	171020	144729	53202	956736
Teens/50s	102152	84576	30724	577130

Table 15. Number of n-grams Generated for each SVM Classification Task (No Stop Words or Entropy-based Words Removed, *Hapax Legomena* Removed).

For the NBC, n-gram dictionaries of both word and character grams were generated for each age group. No stop n-grams, however, were removed from any dictionary. Depending on the experiment feature set, either an empty string replaced punctuation marks, or punctuation marks were retained in the same

Category	Unigrams	Bigrams	Trigrams
Teens	100444	100909	101870
20s	247423	248112	251355
30s	132638	132897	134022
40s	87483	87718	89622
50s	13722	13813	14016
Adult	512376	513639	520418
	3 Character Grams	4 Character Grams	5 Character Grams
Teens	527556	527091	526626
20s	1328342	1327653	1326964
30s	714793	714534	714275
40s	479332	479097	478862
50s	82063	81983	81903
Adult	2775440	2774177	2772914

Table 16. Number of N-grams Generated for Each Bayesian Model.

2. Features for each Classifier

We performed initial experiments using the NBC and a greater number of experiments with the SVM.

a. Naïve Bayes Classifier Features

- Unigrams/bigrams/trigrams with punctuation and no n-grams removed
- Unigrams/bigrams/trigrams without punctuation and no n-grams removed
- 3/4/5 *character* grams

b. Support Vector Machine Features

- Unigrams/bigrams/trigrams with no n-grams removed
- Unigrams/bigrams/trigrams with entropy-based stop n-grams removed
- Unigrams/bigrams/trigrams with mutual high-frequency stop n-grams removed
- Unigrams/bigrams/trigrams with meta-data features
- 3 *character* grams
- Lin Feature Set

3. Naïve Bayes Classifier Setup

Using the training set data, we created two n-gram dictionaries with token counts for each age group. The token counts were used to calculate the conditional probabilities for the NBC. One n-gram dictionary kept punctuation and the other replaced all punctuation with the empty string. We conducted all experiments on both dictionaries. We used Witten-Bell and Laplace Smoothing to assign probability mass to the zero count events, where n-grams appeared in the test set, but not in the training set. For each set of experiments, we applied one type of smoothing.

4. Support Vector Machine Setup

We used the LIBLINEAR [19] library[6] to generate a series of SVM models for each classification task. The models used a linear kernel. Each model was assigned a slack variable using powers of 2 ranging from 2^{-15} to 2^{15}. To generate the vectors for the LIBLINEAR program, we used the n-gram dictionary for each classification task (e.g., teens versus 20s). The n-gram dictionary represented a vector of all the n-gram types found in both age groups in the training set. In other words, each n-gram in the dictionary was mapped to a dimension in the vector.

The dictionary vector represented each document in the training and test sets. The value for each dimension was the frequency of the corresponding n-gram type in the document. As each document vector was generated, the age group of the author was included as a label for that vector. The test document vectors also included the age group, but that label was only used for accounting purposes after the prediction was made. N-grams that occurred in the test set but not in the training set were ignored. For each classification task, training

[6] Freely available from LIBLINEAR—A Library for Large Linear Classification at http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

vectors for both age groups are submitted into the SVM library. The SVM library then generated a model trained for each age group.

For each classification task's model, a different slack variable was applied to each model generated. Therefore, if the task was to generate a model for teens and 20-year-olds using unigram features, we generated 31 different models with 31 different slack variables for that one classification task. Once the SVM library generated a model, the test vectors were submitted into that model and the model predicted the age group for each test vector.

Rather than use a model that used both the test and training data to generate the entire possible vocabulary of each age group, the models in this study only used the training data to generate the vocabulary. This was done to be more representative of chat, where it is likely new words are invented before models can be updated with new vocabulary.

As an example of vector generation, let the training set contain the following vocabulary: {the *yellow dog chased the yellow cat}*. The n-grams in the vocabulary with their frequencies would be mapped to the following unigram dictionary with indexes in brackets: { [0]*cat-1,* [1]*chased-1,* [2]*the-2,* [3]*yellow-2* }. Let a test document contain the following n-grams: *the yellow cat ran into the yellow house.* The vector generated for that document would be [1, 0, 2, 2]. Dimension 1 of the vector represents *cat*, which appears once in the document. Dimension 2 represents *chased*, which does not appear in the document. Dimension 3 represents *the*, which appears twice, and dimension 4 represents *yellow*, which appears twice in the document. Because *house*, *into*, and *ran* are not in the vocabulary dictionary, they do not appear in the vector representing the test document.

5. Random Trials

To test whether or not the models generated were over fitting the data, 10 random trials were conducted. In this research, the test set was chosen

randomly from 20% of the data set. This random sub-sampling from the entire data set was repeated nine additional times. For each training/test set, all the experiments were performed. The average of the F-scores was used to rank and evaluate each feature set for the classification tasks. Rather than use all 10 F-scores generated, the average F-score used all the F-scores except the F-scores with the highest and lowest value. This was done so the average was not as affected by outliers.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS AND ANALYSIS

A. RESULTS

Both the SVM and NBC models demonstrate that they have the capability to distinguish an author's age group. The SVM model performs slightly better than the NBC in all but the teens versus 50s classification task, where the NBC is superior. Also, the SVM model produces significantly better results when classifying teens versus adults. The NBC performs comparably when classifying teens versus specific age groups. Tables 17–20 contain the NBC results and Tables 21–25 contain the SVM results. The tables rank the feature sets by average F-score. In order to exclude outliers, the average F-score was calculated without the highest and lowest F-score measure from the 10 random trials. The tables contain the omitted best and worst F-score results from the 10 random trials. Appendix D contains the detailed results for the NBC. Appendix E contains the detailed results for the SVM.

The SVM results in Tables 21–25 show the results from the model with the slack variable that produced the best average F-score for that feature set. The value of the slack variable can be found in Appendix E. In the SVM results tables, the left column represents the features used in an experiment and if there were any stop n-grams removed or additional meta-data. The label "unigram (5 entropy)" represents the experiment where the features were unigrams and five entropy-based stop n-grams were removed.

Classification Task	Feature	Precision	Recall	F-score Low	F-score High	F-score w/o high/low
Teens vs. 20s	trigram	0.567	0.898	0.438	0.766	0.717
	3 character gram	0.439	0.497	0.054	0.818	0.466
	unigram	0.482	0.461	0.105	0.827	0.462
	4 character gram	0.488	0.411	0.026	0.866	0.433
	5 character gram	0.390	0.238	0.000	0.629	0.285
	bigram	0.278	0.174	0.038	0.419	0.207
Teens vs. 30s	3 character gram	0.804	0.994	0.879	0.899	0.889
	bigram	0.790	0.947	0.627	0.903	0.884
	4 character gram	0.825	0.932	0.774	0.924	0.880
	5 character gram	0.830	0.887	0.561	0.947	0.879
	unigram	0.783	0.953	0.707	0.896	0.873
	trigram	0.756	0.989	0.844	0.869	0.857
Teens vs. 40s	4 character gram	0.921	0.986	0.921	0.970	0.954
	5 character gram	0.916	0.985	0.904	0.967	0.953
	3 character gram	0.901	0.989	0.915	0.963	0.944
	unigram	0.903	0.984	0.890	0.979	0.944
	bigram	0.884	0.993	0.905	0.954	0.937
	trigram	0.877	0.991	0.900	0.963	0.930
Teens vs. 50s	trigram	0.961	0.988	0.957	0.987	0.975
	3 character gram	0.911	1.000	0.939	0.967	0.953
	unigram	0.904	0.999	0.939	0.955	0.950
	bigram	0.899	0.999	0.935	0.963	0.945
	4 character gram	0.895	1.000	0.932	0.959	0.945
	5 character gram	0.881	0.999	0.924	0.947	0.936
Teens vs. Adults	trigram	0.406	0.684	0.245	0.699	0.516
	3 character gram	0.364	0.381	0.022	0.730	0.363
	unigram	0.365	0.304	0.000	0.814	0.305
	4 character gram	0.269	0.190	0.000	0.600	0.199
	bigram	0.205	0.134	0.000	0.347	0.157
	5 character gram	0.170	0.066	0.000	0.275	0.083

* Average F-score computed without highest and lowest F-score

Table 17. Results for NBC with Witten-Bell Smoothing without Punctuation (Ranked Average by F-score for each Classification Task).

Classification Task	Feature	Precision	Recall	F-score Low	F-score High	Average F-score*
Teens vs. 20s	trigram	0.559	0.903	0.194	0.764	0.741
	3 character gram	0.424	0.483	0.011	0.824	0.452
	4 character gram	0.462	0.381	0.025	0.871	0.394
	5 character gram	0.398	0.269	0.000	0.752	0.295
	unigram	0.335	0.236	0.035	0.611	0.261
	bigram	0.283	0.172	0.061	0.383	0.210
Teens vs. 30s	5 character gram	0.834	0.901	0.564	0.939	0.889
	4 character gram	0.829	0.941	0.771	0.921	0.889
	3 character gram	0.804	0.977	0.850	0.903	0.883
	bigram	0.765	0.940	0.531	0.885	0.873
	unigram	0.780	0.932	0.587	0.899	0.872
	trigram	0.749	0.991	0.843	0.867	0.853
Teens vs. 40s	5 character gram	0.922	0.985	0.904	0.971	0.956
	4 character gram	0.912	0.985	0.907	0.967	0.949
	unigram	0.915	0.984	0.908	0.983	0.948
	3 character gram	0.884	0.984	0.874	0.947	0.936
	bigram	0.870	0.995	0.899	0.950	0.929
	trigram	0.840	0.993	0.887	0.939	0.909
Teens vs. 50s	trigram	0.953	0.988	0.952	0.979	0.971
	unigram	0.901	0.999	0.939	0.955	0.948
	3 character gram	0.898	1.000	0.935	0.955	0.946
	bigram	0.894	0.999	0.935	0.955	0.943
	4 character gram	0.892	1.000	0.932	0.955	0.943
	5 character gram	0.888	1.000	0.928	0.951	0.941
Teens vs. Adults	trigram	0.452	0.850	0.141	0.712	0.630
	3 character gram	0.371	0.398	0.000	0.765	0.376
	4 character gram	0.316	0.259	0.000	0.744	0.251
	bigram	0.238	0.159	0.021	0.366	0.187
	5 character gram	0.277	0.135	0.000	0.610	0.145
	unigram	0.201	0.124	0.000	0.376	0.143

* Average F-score computed without highest and lowest F-score

Table 18. Results for NBC with Witten-Bell Smoothing with Punctuation (Ranked by Average F-score for each Classification Task).

Classification Task	Feature	Precision	Recall	F-score Low	F-score High	Average F-score*
Teens vs. 20s	trigram	0.182	0.029	0.000	0.127	0.047
	bigram	0.079	0.003	0.000	0.033	0.004
	3 character gram	0.000	0.000	0.000	0.000	0.000
	4 character gram	0.000	0.000	0.000	0.000	0.000
	5 character gram	0.000	0.000	0.000	0.000	0.000
	unigram	0.000	0.000	0.000	0.000	0.000
Teens vs. 30s	trigram	0.786	0.907	0.440	0.910	0.875
	bigram	0.730	0.709	0.146	0.891	0.744
	unigram	0.475	0.230	0.000	0.901	0.235
	5 character gram	0.507	0.204	0.000	0.935	0.205
	4 character gram	0.385	0.144	0.000	0.922	0.097
	3 character gram	0.167	0.067	0.000	0.702	0.023
Teens vs. 40s	5 character gram	0.926	0.902	0.405	0.979	0.949
	unigram	0.900	0.906	0.338	0.971	0.943
	4 character gram	0.905	0.853	0.081	0.983	0.926
	bigram	0.844	0.997	0.898	0.928	0.914
	trigram	0.817	0.993	0.880	0.913	0.896
	3 character gram	0.873	0.759	0.017	0.983	0.814
Teens vs. 50s	5 character gram	0.859	1.000	0.921	0.932	0.924
	bigram	0.855	1.000	0.921	0.924	0.922
	4 character gram	0.855	1.000	0.921	0.928	0.922
	trigram	0.858	0.995	0.916	0.928	0.921
	3 character gram	0.854	1.000	0.921	0.924	0.921
	unigram	0.854	1.000	0.921	0.924	0.921
Teens vs. Adults	trigram	0.130	0.005	0.000	0.049	0.006
	bigram	0.150	0.002	0.000	0.017	0.002
	3 character gram	0.000	0.000	0.000	0.000	0.000
	4 character gram	0.000	0.000	0.000	0.000	0.000
	5 character gram	0.000	0.000	0.000	0.000	0.000
	unigram	0.000	0.000	0.000	0.000	0.000

* Average F-score computed without highest and lowest F-score

Table 19. Results for NBC with Laplace Smoothing without Punctuation (Ranked by Average F-score for each Classification Task).

Classification Task	Feature	Precision	Recall	F-score Low	F-score High	Average F-score*
Teens vs. 20s	trigram	0.114	0.022	0.000	0.240	0.015
	bigram	0.073	0.003	0.000	0.017	0.004
	3 character gram	0.000	0.000	0.000	0.000	0.000
	4 character gram	0.000	0.000	0.000	0.000	0.000
	5 character gram	0.000	0.000	0.000	0.000	0.000
	unigram	0.000	0.000	0.000	0.000	0.000
Teens vs. 30s	trigram	0.784	0.912	0.520	0.910	0.869
	bigram	0.699	0.665	0.144	0.889	0.691
	5 character gram	0.513	0.164	0.000	0.779	0.183
	unigram	0.406	0.172	0.000	0.839	0.163
	4 character gram	0.369	0.055	0.000	0.325	0.073
	3 character gram	0.185	0.018	0.000	0.167	0.020
Teens vs. 40s	unigram	0.906	0.928	0.548	0.967	0.943
	5 character gram	0.920	0.903	0.475	0.971	0.941
	bigram	0.850	0.997	0.906	0.935	0.916
	trigram	0.827	0.996	0.885	0.921	0.903
	4 character gram	0.852	0.797	0.033	0.975	0.855
	3 character gram	0.864	0.747	0.017	0.979	0.796
Teens vs. 50s	trigram	0.857	0.999	0.916	0.928	0.923
	5 character gram	0.857	1.000	0.921	0.928	0.923
	bigram	0.855	1.000	0.921	0.924	0.922
	4 character gram	0.854	1.000	0.921	0.924	0.921
	3 character gram	0.853	1.000	0.921	0.921	0.921
	unigram	0.854	1.000	0.921	0.924	0.921
Teens vs. Adults	trigram	0.077	0.004	0.000	0.017	0.008
	bigram	0.150	0.002	0.000	0.017	0.002
	3 character gram	0.000	0.000	0.000	0.000	0.000
	4 character gram	0.000	0.000	0.000	0.000	0.000
	5 character gram	0.000	0.000	0.000	0.000	0.000
	unigram	0.000	0.000	0.000	0.000	0.000

* Average F-score computed without highest and lowest F-score

Table 20. Results for NBC with Laplace Smoothing with Punctuation (Ranked by Average F-score for each Classification Task).

Feature Set	Precision	Recall	F-score Low	F-score High	Average F-score*
trigram (25 entropy)	0.694	0.777	0.000	0.987	0.769
trigram (15 entropy)	0.695	0.773	0.000	0.987	0.765
trigram (5 entropy)	0.726	0.753	0.000	0.987	0.760
trigram (50 entropy)	0.685	0.784	0.000	0.987	0.754
trigram	0.675	0.779	0.000	0.987	0.744
trigram (75 entropy)	0.676	0.784	0.000	0.987	0.741
bigram (5 stop)	0.726	0.657	0.000	0.991	0.694
unigram (50 stop)	0.705	0.663	0.000	0.987	0.663
bigram (50 stop)	0.705	0.626	0.000	0.966	0.656
bigram (15 stop)	0.674	0.631	0.000	0.862	0.638
bigram (75 stop)	0.670	0.653	0.000	0.938	0.628
bigram (75 entropy)	0.642	0.648	0.000	0.987	0.625
unigram (25 stop)	0.662	0.607	0.000	0.963	0.615
bigram (50 entropy)	0.619	0.665	0.000	0.987	0.614
bigram	0.614	0.674	0.000	0.987	0.611
trigram (75 stop)	0.691	0.655	0.184	0.897	0.609
bigram (25 entropy)	0.603	0.672	0.000	0.987	0.597
bigram (25 stop)	0.663	0.587	0.000	0.983	0.591
bigram (5 entropy)	0.594	0.674	0.000	0.987	0.585
bigram (15 entropy)	0.594	0.673	0.000	0.987	0.584
trigram (5 stop)	0.681	0.553	0.000	0.970	0.581
trigram (15 stop)	0.586	0.591	0.017	0.970	0.576
unigram (75 stop)	0.775	0.526	0.083	0.825	0.551
unigram (75 entropy)	0.576	0.608	0.099	0.762	0.545
trigram (meta-data)	0.669	0.597	0.033	0.954	0.544
3 character (meta-data)	0.584	0.619	0.014	0.971	0.540
trigram (50 stop)	0.565	0.643	0.065	0.952	0.537
unigram	0.551	0.635	0.099	0.762	0.535
unigram (5 entropy)	0.551	0.634	0.099	0.762	0.534
trigram (25 stop)	0.618	0.533	0.111	0.940	0.524
unigram (25 entropy)	0.540	0.634	0.099	0.762	0.521
bigram (meta-data)	0.602	0.556	0.000	0.946	0.516
unigram (15 stop)	0.625	0.584	0.017	0.963	0.515
unigram (15 entropy)	0.532	0.634	0.099	0.762	0.509
unigram (50 entropy)	0.529	0.634	0.099	0.762	0.505
unigram (5 stop)	0.682	0.526	0.079	0.978	0.502
3 character	0.525	0.626	0.021	0.872	0.494
unigram (meta-data)	0.573	0.539	0.058	0.987	0.483
Lin features	0.420	0.469	0.058	0.558	0.468

* Average F-score computed without highest and lowest F-score

Table 21. Teens vs. 20s SVM Results (Ranked by Average F-score).

Feature Set	Precision	Recall	F-score Low	F-score High	Average F-score*
bigram (15 stop)	0.880	0.945	0.734	0.991	0.914
bigram (50 stop)	0.878	0.960	0.780	1.000	0.913
unigram (50 stop)	0.865	0.904	0.307	1.000	0.894
bigram (25 stop)	0.857	0.957	0.782	1.000	0.893
bigram (5 stop)	0.876	0.924	0.742	0.996	0.891
trigram (5 stop)	0.880	0.896	0.550	0.987	0.888
unigram (15 stop)	0.884	0.858	0.098	1.000	0.888
bigram (75 stop)	0.862	0.944	0.780	1.000	0.888
trigram (15 stop)	0.870	0.909	0.594	0.982	0.888
unigram (25 stop)	0.894	0.832	0.067	0.978	0.886
trigram (25 stop)	0.829	0.968	0.781	0.975	0.885
bigram (50 entropy)	0.819	0.841	0.094	1.000	0.884
unigram (5 stop)	0.807	0.854	0.088	1.000	0.883
trigram (25 entropy)	0.761	0.872	0.000	0.991	0.882
bigram (25 entropy)	0.819	0.838	0.094	1.000	0.881
trigram (15 entropy)	0.761	0.871	0.000	0.991	0.881
unigram (75 stop)	0.841	0.897	0.098	1.000	0.878
unigram (25 entropy)	0.845	0.860	0.462	1.000	0.865
trigram (50 entropy)	0.731	0.879	0.000	0.991	0.863
trigram (5 entropy)	0.733	0.872	0.000	0.991	0.863
trigram	0.727	0.879	0.000	0.991	0.860
trigram (75 entropy)	0.727	0.879	0.000	0.991	0.860
bigram (75 entropy)	0.787	0.841	0.094	1.000	0.860
unigram (75 entropy)	0.837	0.862	0.462	1.000	0.858
unigram (50 entropy)	0.836	0.862	0.462	1.000	0.858
bigram (15 entropy)	0.785	0.841	0.094	1.000	0.858
bigram (5 entropy)	0.785	0.841	0.094	1.000	0.858
unigram (15 entropy)	0.835	0.861	0.462	1.000	0.856
unigram	0.834	0.859	0.462	1.000	0.855
unigram (5 entropy)	0.834	0.859	0.462	1.000	0.855
bigram	0.790	0.828	0.094	1.000	0.852
trigram (75 stop)	0.764	0.818	0.000	0.996	0.852
trigram (50 stop)	0.810	0.870	0.254	1.000	0.850
3 character	0.740	0.822	0.015	1.000	0.839
3 character (meta-data)	0.710	0.822	0.015	1.000	0.816
bigram (meta-data)	0.759	0.785	0.097	1.000	0.808
unigram (meta-data)	0.750	0.822	0.295	0.987	0.794
Lin features	0.649	0.991	0.777	0.789	0.785
trigram (meta-data)	0.750	0.780	0.050	0.946	0.765

* Average F-score computed without highest and lowest F-score

Table 22. Teens vs. 30s SVM Results (Ranked by Average F-score).

Feature Set	Precision	Recall	F-score Low	F-score High	Average F-score*
trigram (75 stop)	0.959	0.995	0.808	1.000	0.991
bigram (75 stop)	0.994	0.902	0.461	1.000	0.980
trigram (25 stop)	0.964	0.969	0.836	1.000	0.977
bigram (50 stop)	0.996	0.893	0.430	1.000	0.976
unigram (75 stop)	0.979	0.936	0.710	1.000	0.975
trigram (15 stop)	0.968	0.951	0.792	1.000	0.974
trigram (50 stop)	0.947	0.984	0.836	1.000	0.974
unigram	0.987	0.893	0.538	1.000	0.962
unigram (5 entropy)	0.987	0.893	0.538	1.000	0.962
unigram (50 stop)	0.989	0.878	0.487	1.000	0.960
unigram (15 entropy)	0.985	0.892	0.538	1.000	0.959
unigram (50 entropy)	0.984	0.892	0.538	1.000	0.959
unigram (75 entropy)	0.984	0.892	0.538	1.000	0.959
bigram (25 stop)	0.965	0.932	0.735	1.000	0.959
trigram	0.950	0.958	0.853	1.000	0.957
trigram (15 entropy)	0.950	0.958	0.853	1.000	0.957
trigram (25 entropy)	0.950	0.958	0.853	1.000	0.957
unigram (25 entropy)	0.987	0.886	0.538	1.000	0.957
trigram (5 entropy)	0.950	0.958	0.856	1.000	0.957
trigram (5 stop)	0.978	0.928	0.833	1.000	0.956
trigram (75 entropy)	0.949	0.957	0.853	1.000	0.956
trigram (50 entropy)	0.950	0.951	0.853	1.000	0.953
bigram (25 entropy)	0.966	0.905	0.594	1.000	0.952
bigram	0.932	0.863	0.094	1.000	0.952
bigram (50 entropy)	0.932	0.863	0.094	1.000	0.952
bigram (15 entropy)	0.965	0.903	0.568	1.000	0.952
bigram (5 entropy)	0.963	0.903	0.568	1.000	0.951
bigram (75 entropy)	0.929	0.863	0.094	1.000	0.950
bigram (15 stop)	0.989	0.870	0.541	1.000	0.947
unigram (25 stop)	0.967	0.798	0.127	1.000	0.910
unigram (15 stop)	0.993	0.798	0.173	1.000	0.910
3 character	0.931	0.832	0.400	1.000	0.907
3 character (meta-data)	0.931	0.832	0.400	1.000	0.907
bigram (5 stop)	0.978	0.837	0.667	1.000	0.902
bigram (meta-data)	0.931	0.820	0.378	0.996	0.896
unigram (meta-data)	0.942	0.803	0.439	1.000	0.877
unigram (5 stop)	0.970	0.772	0.188	1.000	0.877
trigram (meta-data)	0.902	0.792	0.430	1.000	0.846
Lin features	0.728	0.959	0.796	0.867	0.827

* Average F-score computed without highest and lowest F-score

Table 23. Teens vs. 40s SVM Results (Ranked by Average F-score).

Feature Set	Precision	Recall	F-score Low	F-score High	Average F-score*
trigram (50 stop)	0.910	0.966	0.589	0.996	0.958
trigram (25 stop)	0.909	0.959	0.583	0.996	0.953
trigram (75 stop)	0.931	0.884	0.366	1.000	0.951
trigram (15 stop)	0.927	0.951	0.736	0.996	0.951
bigram (25 entropy)	0.923	0.892	0.388	0.996	0.946
unigram (75 stop)	0.943	0.921	0.643	1.000	0.945
bigram (75 stop)	0.955	0.904	0.594	1.000	0.945
trigram (5 stop)	0.914	0.932	0.583	0.996	0.935
3 character (meta-data)	0.934	0.913	0.635	1.000	0.932
trigram	0.975	0.831	0.202	1.000	0.925
Lin features	0.814	0.998	0.589	0.928	0.922
bigram (50 entropy)	0.942	0.832	0.245	0.996	0.921
bigram (15 entropy)	0.889	0.886	0.388	0.996	0.917
unigram (50 entropy)	0.951	0.841	0.333	1.000	0.914
bigram (50 stop)	0.922	0.893	0.627	0.996	0.913
bigram (15 stop)	0.946	0.849	0.487	1.000	0.913
bigram	0.880	0.890	0.388	0.996	0.910
trigram (25 entropy)	0.938	0.834	0.202	1.000	0.903
unigram	0.908	0.891	0.657	0.991	0.900
unigram (5 entropy)	0.908	0.890	0.655	0.991	0.899
unigram (15 entropy)	0.898	0.897	0.590	0.991	0.898
unigram (25 entropy)	0.933	0.851	0.636	0.991	0.898
bigram (5 entropy)	0.912	0.891	0.586	1.000	0.896
trigram (15 entropy)	0.936	0.827	0.202	1.000	0.896
bigram (75 entropy)	0.868	0.887	0.388	0.991	0.895
unigram (50 stop)	0.904	0.891	0.522	0.996	0.894
3 character	0.930	0.845	0.519	0.967	0.892
trigram (5 entropy)	0.914	0.826	0.202	1.000	0.872
bigram (25 stop)	0.919	0.778	0.050	0.996	0.871
trigram (50 entropy)	0.915	0.823	0.202	1.000	0.870
trigram (75 entropy)	0.921	0.814	0.202	1.000	0.865
unigram (25 stop)	0.916	0.787	0.159	0.996	0.848
unigram (5 stop)	0.911	0.792	0.188	0.991	0.847
unigram (meta-data)	0.882	0.747	0.092	0.996	0.823
unigram (75 entropy)	0.886	0.797	0.188	0.996	0.823
unigram (15 stop)	0.872	0.779	0.162	0.996	0.805
bigram (meta-data)	0.887	0.775	0.435	1.000	0.800
trigram (meta-data)	0.899	0.709	0.307	0.983	0.760
bigram (5 stop)	0.867	0.730	0.120	1.000	0.738

* Average F-score computed without highest and lowest F-score

Table 24. Teens vs. 50s SVM Results (Ranked by Average F-score).

Feature Set	Precision	Recall	F-score Low	F-score High	Average F-score*
unigram (5 stop)	0.846	0.772	0.114	0.991	0.786
bigram (15 stop)	0.709	0.821	0.033	0.996	0.766
bigram (5 stop)	0.686	0.821	0.038	1.000	0.757
trigram	0.647	0.870	0.025	1.000	0.756
trigram (5 entropy)	0.644	0.870	0.025	1.000	0.754
trigram (75 entropy)	0.646	0.868	0.026	1.000	0.753
trigram (50 entropy)	0.645	0.856	0.026	1.000	0.745
trigram (5 stop)	0.682	0.855	0.167	0.991	0.743
trigram (15 entropy)	0.646	0.844	0.025	1.000	0.737
unigram (15 stop)	0.742	0.754	0.000	1.000	0.735
trigram (25 entropy)	0.654	0.822	0.000	1.000	0.734
trigram (50 stop)	0.778	0.832	0.294	0.991	0.732
bigram (25 stop)	0.648	0.850	0.031	1.000	0.729
bigram (50 stop)	0.671	0.812	0.014	1.000	0.728
trigram (15 stop)	0.686	0.803	0.044	1.000	0.721
unigram (75 stop)	0.689	0.780	0.000	1.000	0.720
bigram (75 stop)	0.631	0.854	0.000	1.000	0.716
trigram (25 stop)	0.749	0.829	0.294	0.987	0.708
trigram (75 stop)	0.782	0.797	0.294	1.000	0.705
bigram (25 entropy)	0.727	0.828	0.294	1.000	0.700
unigram (25 stop)	0.702	0.733	0.000	1.000	0.698
bigram (50 entropy)	0.723	0.832	0.294	1.000	0.698
bigram (5 entropy)	0.722	0.828	0.294	1.000	0.694
bigram	0.716	0.834	0.294	1.000	0.692
bigram (75 entropy)	0.711	0.828	0.294	1.000	0.688
bigram (15 entropy)	0.710	0.828	0.294	1.000	0.686
unigram (50 stop)	0.683	0.737	0.000	0.996	0.679
unigram (50 entropy)	0.713	0.733	0.174	1.000	0.670
unigram (5 entropy)	0.707	0.730	0.138	0.996	0.670
unigram (15 entropy)	0.702	0.728	0.112	0.996	0.670
unigram	0.706	0.729	0.138	0.996	0.669
unigram (25 entropy)	0.702	0.714	0.112	0.996	0.660
unigram (75 entropy)	0.692	0.728	0.138	0.996	0.659
unigram (meta-data)	0.680	0.701	0.067	1.000	0.640
trigram (meta-data)	0.540	0.785	0.029	0.996	0.640
bigram (meta-data)	0.651	0.740	0.191	0.996	0.584
3 character (meta-data)	0.633	0.627	0.067	1.000	0.543
3 character	0.622	0.627	0.067	0.960	0.541
Lin features	0.432	0.361	0.199	0.921	0.327

* Average F-score computed without highest and lowest F-score

Table 25. Teens vs. Adults SVM Results (Ranked by Average F-score).

B. ANALYSIS

The following section contains analysis of the NBC's and SVM's performance. As part of the SVM discussion, the two different types of stop n-gram lists are compared and the addition of meta-data is discussed. When analyzing the effect of stop n-grams and meta-data, we compare the results from the base n-gram feature (no n-grams removed or addition of meta-data) to the results from the n-gram feature with stop n-grams or meta-data. For example, the result from the "unigram (5 entropy)" feature set experiment is compared to "unigram" experiment. As a follow-up to Lin's recommendation to use a SVM, this section also contains a comparison between the SVM and NBC, using her features [5].

1. Naïve Bayes Classifier

The NBC model with Witten-Bell smoothing performs better than the model with Laplace smoothing. One of the weaknesses of Laplace smoothing manifests when data is sparse; too much probability mass is given to the zero count events. With more than 85,000 types in the vocabulary and an average of 339.2 tokens per author, the data for this experiment is very sparse.

The NBC most likely does not do as well when classifying teens against adults, because of the unbalanced data set. Unbalanced data sets cause undue influence of the prior probability. The prior probability favors adults because of the greater number of adult authors in the training set. In the training set, more than 73% of the authors are adults (Table 13). Also, the distribution of n-grams favor adults as well, where over 80% of n-gram types in the training set were written by adults (Table 16).

Using style-based features, Kucukyilmaz et al. had an accuracy result of 0.754, when comparing people age 24 and under to people older than 24 [11]. That result is better than this research's best NBC accuracy result (0.698) comparing teens against adults. Term-based results (unigram) in this research

generated a better accuracy (0.696) than the unigram NBC results (0.608) [11] of Kucukyilmaz et al. Term-based features were more distinguishing than style-based features in the data set used for this research. There may be differences between term/style features in the Turkish language, compared to the English language, making style-based features more distinctive than term-based features.

The results, however, may not be directly comparable, because our experiments did not separate authors at the age of 24. Because of the closeness in age to teens, including authors of age 20–24 in the younger age group, this may have made the classification task easier.

Another difference in the setup of experiments, Kucukyilmaz et al. used categories defined by thresholds as values for their features (e.g., low/medium/high, exist/not exist, short/average/long, rich/poor/average) instead of actual style frequency counts [11]. By having such values, they did not have to deal with sparse vectors, which may have led to the better results for their NBC with Laplace smoothing. They also balanced their data set, so they did not have the problem of an unbalanced data set, where the prior probability can have caused undue influence. Future experiments using a balanced data set, term-based features, and collapsing the frequency count to categories for style-based features could improve performance of the NBC.

When classifying teens against 20s, 30s, 40s, and 50s, higher order n-grams, to include *character* grams, performed better than Lin's feature set of punctuation marks and emoticons. Given the better results, especially when classifying teens against 20-year-olds, it appears that context is a necessary feature. Lin had compared people age 25 and under against people older than 25 (0.702 F-score) [11], so one cannot directly compare the best result from the NBC teens versus adults classification in this experiment (0.630 F-score).

Arguably, a teen's vocabulary is more likely to be similar to people in their early 20s, so the teen versus adult classification task in this experiment is more difficult.

2. Support Vector Machine

The SVM model performs slightly better than the NBC in all age group classification tasks, except in the teens versus 50s task. It performs significantly better when classifying teens versus adults. The SVM is better able to overcome the unbalanced data problem with that classification task. There is, however, great variation in F-score results between the different random trials, especially in the teens versus 20s classification task.

A possible reason for the variation could be that the training set for the 20-year-olds contains a disproportionate distribution of authors within the age group. Thus, if there were not enough text written by people in their early 20s in the training set, the SVM model would be poorly trained, and could misidentify authors in their early 20s as teenagers. Table 26 contains the distribution of the authors in the 20-year-old age group, for each random trial's training set. Table 27 contains the combined file sizes (bytes) of authors of age 20 and 21; 20 to 24; and 25 to 29 in each random training set. Sets 7 and 8 are in bold because those two sets caused the worst performance in the SVM models.

Age	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	**Set 7**	**Set 8**	Set 9	Set 10
20	80	78	77	82	76	73	**79**	**78**	78	78
21	98	96	101	93	94	97	**96**	**97**	101	99
22	82	81	84	86	85	84	**86**	**83**	83	78
23	65	68	66	67	65	70	**71**	**66**	70	69
24	61	62	65	67	64	64	**60**	**66**	61	64
25	80	84	80	77	82	80	**79**	**81**	71	83
26	69	72	65	67	75	67	**70**	**66**	65	64
27	65	63	61	61	65	58	**56**	**60**	65	68
28	57	51	52	55	51	56	**56**	**56**	57	53
29	32	34	38	34	32	40	**36**	**36**	38	33

Table 26. Distribution of Authors in the 20s Age Group per Random Training Data Set.

Age	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	**Set 7**	**Set 8**	Set 9	Set 10
20-21	229118	221011	229272	240508	188378	232550	**203907**	**248338**	201428	215699
20-24	704722	751978	607157	707521	726645	616140	**727175**	**713301**	715451	671532
25-29	623620	559732	624152	601768	714411	602779	**545940**	**616241**	676997	548478
20-29	1328342	1311710	1231309	1309289	1441056	1218919	**1273115**	**1329542**	1392448	1220010

Table 27. Combined File Sizes (Bytes) of Authors of Ages 20 to 21, 20 to 24 and 25 to 29 per Random Training Set.

Based on a visual inspection of Table 26, there did not seem to be a disproportionate distribution of authors in the different random training sets. Table 27 does not indicate that there is a correlation between the amount of text per age group and the poor performance of training sets 7 and 8.

Tables 28 and 29 contain the distribution of the teen authors in the teen age group and the combined file sizes (bytes) of ages 13 to 17; 18 and 19; and 13 to 19. The distribution of files by 18- and 19-year-olds is shown, because those authors are more likely to be similar to 20-year-olds. Again, training sets 7 and 8 are highlighted, because they were the worst performing data sets.

Age	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	**Set 7**	**Set 8**	Set 9	Set 10
13	7	4	6	7	6	5	**6**	**5**	5	6
14	43	37	38	42	42	38	**37**	**35**	30	38
15	51	47	52	48	47	56	**51**	**49**	56	52
16	79	86	80	83	77	87	**83**	**82**	82	79
17	60	67	70	72	72	69	**76**	**76**	66	66
18	114	107	116	109	109	104	**104**	**115**	110	112
19	111	117	103	104	112	106	**108**	**103**	116	112

Table 28. Distribution of Teen Authors in Random Training Sets.

Age	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	**Set 7**	**Set 8**	Set 9	Set 10
13-17	218834	196569	230844	238519	206102	235757	**212591**	**199877**	210643	213696
18-19	308722	266865	292824	212081	297161	312138	**318179**	**218427**	247682	330006
13-19	527556	463434	523668	450600	503263	547895	**530770**	**418304**	458325	543702

Table 29. Combined File Sizes (Bytes) of Authors of Ages 13 to 17, 18 to 19, and 13–19 per Random Training Set.

Similar to the 20-year-olds, the distribution of ages in the random training data sets is nearly balanced. The combined file size of training set 8, however, is significantly less than the other training sets. In set 8, there may not be enough teen training data to adequately train the SVM model, thus causing the extremely poor performance in the teens versus 20s classification task. The size of training set 7 is similar to the other training sets, so in this case, the n-grams themselves may be the cause of the poor performance and not the amount of text available to train the model.

Compared to the SVM model Kucukyilmaz et al. generated, the SVM model in this research has an improved accuracy result, when comparing teens against adults using unigrams as features. Their accuracy was 0.563 [11] and the SVM model in our experiments had an accuracy of 0.780 when using unigrams as features. Higher order n-grams perform even better. The results may not be directly comparable because our experiments did not include people age 24 and under in the younger age data set.

We believe that including people in their early 20s, as part of the teen data set, makes the classification task easier, because of the similarity of conversations by teens and people in their early 20s. To verify this, we relaxed the definition of a teen to be someone between the ages of 13–21, because some 20- and 21-year-olds may converse more like a teen. Similarly, we relaxed the definition of the 20s age group to be someone between the ages of 18–29. Some 18- and 19-year-olds may be more mature and have more adult-like conversations. In the teens versus 20s classification task, using the relaxed age groups improved the average F-score by 0.121. Table 30 shows the F-score results for each feature set in the teens versus 20s classification task, using the strict definition of the 20s age group (20–29) and the relaxed definition of the teens/20s age group.

Feature Set	Strict		Relaxed		Strict	Relaxed	Change in
	Precision	Recall	Precision	Recall	F-score	F-score	F-score
Trigram (50 entropy)	0.685	0.784	0.796	0.833	0.754	0.845	0.091
Trigram (75 entropy)	0.676	0.784	0.795	0.833	0.741	0.844	0.103
Trigram (25 entropy)	0.694	0.777	0.789	0.833	0.769	0.841	0.072
Trigram (15 entropy)	0.695	0.773	0.790	0.828	0.765	0.838	0.073
Trigram (5 entropy)	0.726	0.753	0.789	0.828	0.760	0.838	0.078
Trigram	0.675	0.779	0.789	0.828	0.744	0.838	0.093
Unigram (50 mutual)	0.705	0.663	0.805	0.721	0.663	0.783	0.120
Bigram (5 mutual)	0.726	0.657	0.776	0.732	0.694	0.780	0.086
Bigram (25 mutual)	0.663	0.587	0.800	0.734	0.591	0.770	0.179
Bigram (15 mutual)	0.674	0.631	0.806	0.705	0.638	0.750	0.112
Trigram (75 mutual)	0.691	0.655	0.774	0.769	0.609	0.740	0.131
Bigram (50 mutual)	0.705	0.626	0.798	0.698	0.656	0.733	0.078
Unigram (75 mutual)	0.775	0.526	0.884	0.636	0.551	0.725	0.174
Bigram (75 mutual)	0.670	0.653	0.751	0.706	0.628	0.724	0.096
Unigram (25 mutual)	0.662	0.607	0.771	0.659	0.615	0.718	0.103
Trigram (5 mutual)	0.681	0.553	0.753	0.681	0.581	0.701	0.120
Bigram (75 entropy)	0.642	0.648	0.698	0.714	0.625	0.697	0.073
Bigram	0.614	0.674	0.698	0.714	0.611	0.697	0.086
Bigram (5 entropy)	0.594	0.674	0.698	0.714	0.585	0.697	0.112
Bigram (50 entropy)	0.619	0.665	0.698	0.713	0.614	0.697	0.083
Bigram (25 entropy)	0.603	0.672	0.698	0.713	0.597	0.696	0.099
Bigram (15 entropy)	0.594	0.673	0.698	0.713	0.584	0.696	0.113
Trigram (15 mutual)	0.586	0.591	0.702	0.705	0.576	0.694	0.119
Trigram (50 mutual)	0.565	0.643	0.697	0.755	0.537	0.693	0.156
Unigram	0.551	0.635	0.711	0.696	0.535	0.681	0.146
Unigram (5 entropy)	0.551	0.634	0.703	0.696	0.534	0.681	0.147
Unigram (25 entropy)	0.540	0.634	0.670	0.717	0.521	0.680	0.159
Unigram (15 entropy)	0.532	0.634	0.670	0.717	0.509	0.680	0.171
Unigram (75 entropy)	0.576	0.608	0.668	0.717	0.545	0.678	0.133
Unigram (50 entropy)	0.529	0.634	0.668	0.717	0.505	0.678	0.173
Trigram (meta-data)	0.669	0.597	0.763	0.679	0.544	0.668	0.123
Unigram (15 mutual)	0.625	0.584	0.754	0.630	0.515	0.660	0.145
3 Character (meta-data)	0.584	0.619	0.680	0.720	0.540	0.657	0.117
3 Character	0.525	0.626	0.680	0.720	0.494	0.657	0.163
Unigram (5 mutual)	0.682	0.526	0.778	0.607	0.502	0.655	0.153
Lin Features	0.420	0.469	0.608	0.728	0.468	0.652	0.184
Trigram (25 mutual)	0.618	0.533	0.688	0.663	0.524	0.642	0.117
Bigram (meta-data)	0.602	0.556	0.684	0.650	0.516	0.617	0.100
Unigram (meta-data)	0.573	0.539	0.681	0.638	0.483	0.615	0.132

Table 30. Comparison of Performance when Classifying Teens Versus 20s Using Strict and Relaxed Teen Age Groups (Ranked by Relaxed F-score).

Because there is such similarity in the conversations, both age groups may occupy the same vector space area, so a linear kernel may not produce the best separation even with different slack variables. Given the success of the SVM with the 30s, 40s, and 50s age group, it demonstrated that is has the capability of distinguishing age groups. Future experiments using a different kernel type (e.g., polynomial or radial), may generate better results for the teens versus 20s/adult classification task.

3. Entropy-Based Stop Words

The entropy-based stop words generated were dissimilar from the mutual high-frequency stop words generated. Table 31 contains the first 10 high-frequency-based stop words and entropy-based stop words generated from the first random training set. These stop words were used to classify teens versus adults with unigrams as the feature. In general, there were only a few cases where an intersection occurred between the entropy-based lists and the high-frequency-based lists. Table 32 contains the set of the intersections that occurred in the random training data sets. In some data sets, some, but not all, the n-grams listed intersected.

High-Frequency n-gram	Token Count
i	15318
lol	12740
the	9768
a	9260
to	9258
you	7847
and	6366
is	6115
.action	5566
hi	5137

Entropy n-gram	Token Count
hoho.	2
hops	4
late.	4
overweight	2
hustla	18
pie.	2
object	2
dosent	3
scott,	2
ours	2

Table 31. Comparison of High-Frequency and Entropy-Based Stop Unigrams and Their Usage.

Age Classification	Unigrams	Bigrams	Trigrams
Teens/30s	I up you		<post> im a <post> im not
Teens/40s	is up what yes	I can on the	I want to have a good <post> who is <post> what are <post> ;-) </post>
Teens/50s	go hi not too yes		<post> any ladies <post> lol @ <post> what is

Table 32. Intersection of High-Frequency and Entropy-Based Stop N-grams.

In almost all stop n-gram lists, the entropy of the n-gram was one—an even distribution of the n-grams between the teen and older age group class. There were three cases where an n-gram's entropy value was less than one. In the Unigram (75 entropy) features set, there were unigrams with an entropy of 0.985 that were removed. In the Trigram (50 entropy) and the Trigram (75 entropy) feature sets, there were trigrams removed with an entropy value as low as 0.971 and 0.896 respectively. In those three instances, the removal of n-grams with entropy lower than 1.000 caused greater degradation to the F-score, compared to the other cases where only n-grams with entropy of one were removed. Table 33 shows the difference in F-score from the feature sets that had entropy-based n-grams removed to the feature sets that had no n-grams removed. In the table, the F-score is the average F-score calculated without the highest and lowest F-score measure from the 10 random trials. Table 33 also bolds the experiment results where n-gram entropy values were less than 1.000. Appendix B contains the lists of the 75 entropy-based stop n-grams generated for the seventh random data set, which contained n-grams with entropy values of less than 1.000.

Classification Task/ Feature Set	Teens vs. 20s		Teens vs. 30s		Teens vs. 40s		Teens vs. 50s		Teens vs. Adults	
	F-score	Difference	F-score	Difference	F-score	Difference	F-score	Difference	F-score	Difference
Unigram	0.535	0.000	0.855	0.000	0.962	0.000	0.900	0.000	0.669	0.000
Unigram (5 entropy)	0.534	-0.001	0.855	0.000	0.962	0.000	0.899	-0.001	0.670	0.001
Unigram (15 entropy)	0.509	-0.026	0.856	0.001	0.959	-0.002	0.898	-0.002	0.670	0.001
Unigram (25 entropy)	0.521	-0.015	0.865	0.010	0.957	-0.005	0.898	-0.002	0.660	-0.008
Unigram (50 entropy)	0.505	-0.031	0.858	0.003	0.959	-0.003	0.914	0.013	0.670	0.001
Unigram (75 entropy)	0.545	0.010	0.858	0.003	0.959	-0.003	**0.823**	**-0.078**	0.659	-0.010
Bigram	0.611	0.000	0.852	0.000	0.952	0.000	0.910	0.000	0.692	0.000
Bigram (5 entropy)	0.585	-0.026	0.858	0.005	0.951	-0.001	0.896	-0.014	0.694	0.002
Bigram (15 entropy)	0.584	-0.028	0.858	0.005	0.952	0.000	0.917	0.008	0.686	-0.006
Bigram (25 entropy)	0.597	-0.014	0.881	0.029	0.952	0.000	0.946	0.036	0.700	0.008
Bigram (50 entropy)	0.614	0.002	0.884	0.032	0.952	0.000	0.921	0.011	0.698	0.006
Bigram (75 entropy)	0.625	0.014	0.860	0.007	0.950	-0.002	0.895	-0.015	0.688	-0.004
Trigram	0.744	0.000	0.860	0.000	0.957	0.000	0.925	0.000	0.756	0.000
Trigram (5 entropy)	0.760	0.016	0.863	0.002	0.957	0.000	0.872	-0.053	0.754	-0.002
Trigram (15 entropy)	0.765	0.021	0.881	0.021	0.957	0.000	0.896	-0.029	0.737	-0.019
Trigram (25 entropy)	0.769	0.025	0.882	0.021	0.957	0.000	0.903	-0.022	0.734	-0.022
Trigram (50 entropy)	0.754	0.010	0.863	0.003	0.953	-0.004	**0.870**	**-0.055**	0.745	-0.012
Trigram (75 entropy)	0.741	-0.003	0.860	0.000	0.956	-0.002	**0.865**	**-0.060**	0.753	-0.003

Table 33. Effect on F-score as Increasing Number of Entropy-Based Stop N-grams are Removed.

The experiments with the removal of entropy-based n-grams generated slight gains and losses in performance. The highest gain was 0.036. The largest losses were -0.078, -0.055, and -0.060. Those losses occurred when feature sets removed stop n-grams with an entropy value of less than 1.000. If the values from those experiments are excluded, then the greatest degradation in F-score was 0.053 from the base case (no n-grams removed). Without the feature sets where entropy values were less than 1.000, the overall average gain is 0.008 and the overall average loss is -0.012. Those averages indicate that removal of entropy-based n-grams with entropy values of 1.000 may only have a slight impact on performance.

The differences in gain/loss do not seem to be monotonic, as more n-grams are removed. In this research, entropy values were calculated across an age group, rather than normalized by document. That could be the reason why a monotonic increase/decrease is not seen. Future experiments with entropy values normalized for each document could better determine the effects of

removing an increasing number of entropy-based stop n-grams. Those experiments could also explore the effects of different entropy thresholds, to determine to what point it is worthwhile to remove n-grams that have entropy values of less than 1.000.

4. Mutual High-Frequency Stop Words

When comparing the difference between the feature set with no n-grams removed and the feature set with mutual high-frequency stop words removed, the removal of the n-grams had more of an effect in performance, compared to the entropy-based experiments. The highest gain was 0.128 and largest loss was -0.220. The overall average gain was 0.038 and loss was -0.065. Appendix C contains the lists of the 75 high-frequency-based stop words generated for the sixth random training set, one of the better performing data sets.

Table 34 shows the difference in F-score from the feature sets that had n-grams removed to the feature sets that had no n-grams removed. In the table, the F-score is the average F-score calculated without the highest and lowest F-score measure from the 10 random trials. Similar to the behavior of the entropy-based stop words, there does not seem to be a monotonic increase/decrease in performance as more n-grams are removed. In this research, the frequency counts of the n-grams were calculated across an age group, which may have caused the non-monotonic increase/decrease in performance. Future experiments with frequency values normalized for each document could better determine the effects of removing more and more stop n-grams. Such an approach will be equivalent to using tf-idf for feature selection to reduce dimensionality.

Classification Task/ Feature Set	Teens vs. 20s		Teens vs. 30s		Teens vs. 40s		Teens vs. 50s		Teens vs. Adults	
	F-score	Difference	F-score	Difference	F-score	Difference	F-score	Difference	F-score	Difference
Unigram	0.535	0.000	0.855	0.000	0.962	0.000	0.900	0.000	0.669	0.000
Unigram (5 stop)	0.502	-0.033	0.883	0.028	0.877	-0.085	0.847	-0.053	0.786	0.117
Unigram (15 stop)	0.515	-0.020	0.888	0.034	0.910	-0.052	0.805	-0.095	0.735	0.066
Unigram (25 stop)	0.615	0.080	0.886	0.031	0.910	-0.051	0.848	-0.052	0.698	0.029
Unigram (50 stop)	0.663	0.128	0.894	0.039	0.960	-0.002	0.894	-0.006	0.679	0.010
Unigram (75 stop)	0.551	0.016	0.878	0.024	0.975	0.014	0.945	0.045	0.720	0.051
Bigram	0.611	0.000	0.852	0.000	0.952	0.000	0.910	0.000	0.692	0.000
Bigram (5 stop)	0.694	0.083	0.891	0.039	0.902	-0.050	0.738	-0.171	0.757	0.065
Bigram (15 stop)	0.638	0.027	0.914	0.062	0.947	-0.005	0.913	0.003	0.766	0.074
Bigram (25 stop)	0.591	-0.020	0.893	0.041	0.959	0.006	0.871	-0.039	0.729	0.038
Bigram (50 stop)	0.656	0.045	0.913	0.060	0.976	0.024	0.913	0.003	0.728	0.037
Bigram (75 stop)	0.628	0.017	0.888	0.036	0.980	0.027	0.945	0.036	0.716	0.024
Trigram	0.744	0.000	0.860	0.000	0.957	0.000	0.925	0.000	0.756	0.000
Trigram (5 stop)	0.581	-0.163	0.888	0.028	0.956	-0.001	0.935	0.010	0.743	-0.013
Trigram (15 stop)	0.576	-0.168	0.888	0.028	0.974	0.017	0.951	0.026	0.721	-0.035
Trigram (25 stop)	0.524	-0.220	0.885	0.024	0.977	0.020	0.953	0.028	0.708	-0.049
Trigram (50 stop)	0.537	-0.207	0.850	-0.011	0.974	0.017	0.958	0.033	0.732	-0.024
Trigram (75 stop)	0.609	-0.135	0.852	-0.008	0.991	0.034	0.951	0.026	0.705	-0.051

Table 34. Effect on F-score as Increasing Number of High-Frequency-Based Stop N-grams are Removed.

There was a significant decrease in performance when stop words were removed from the trigram feature. A possible reason is that the n-grams removed were very distinct for one age group, where one age group used some n-grams more frequently than the other age group. As an example, in the first random training data set, the trigram *<beginning of post tag> .action is* was removed in all feature sets that use stop word removal. Teens wrote that trigram 326 times. The 20s usage was 708; 30s was 202; 40s was 65; and 50s was 5. Teens only used that n-gram 5% of the time compared to adults. By removing that n-gram, that distinguishing n-gram was lost.

Table 35 shows the average percentage of use by teens and adults of the removed words for each age group classification task.

	Unigram Average		Bigram Average		Trigram Average	
Mutual Words Removed	Teens	20s	Teens	20s	Teens	20s
5	0.280	0.734	0.287	0.713	0.168	0.832
15	0.286	0.718	0.273	0.727	0.221	0.779
25	0.283	0.720	0.279	0.721	0.240	0.760
50	0.278	0.723	0.275	0.725	0.263	0.737
75	0.279	0.721	0.278	0.722	0.269	0.731
Mutual Words Removed	Teens	30s	Teens	30s	Teens	30s
5	0.400	0.600	0.439	0.561	0.354	0.646
15	0.396	0.604	0.412	0.588	0.386	0.614
25	0.382	0.618	0.424	0.576	0.407	0.593
50	0.393	0.607	0.410	0.590	0.436	0.564
75	0.388	0.612	0.410	0.590	0.425	0.575
Mutual Words Removed	Teens	40s	Teens	40s	Teens	40s
5	0.454	0.546	0.497	0.503	0.402	0.598
15	0.468	0.532	0.441	0.559	0.494	0.506
25	0.467	0.533	0.470	0.530	0.497	0.503
50	0.471	0.529	0.474	0.526	0.519	0.481
75	0.474	0.526	0.486	0.514	0.509	0.491
Mutual Words Removed	Teens	50s	Teens	50s	Teens	50s
5	0.839	0.161	0.856	0.144	0.808	0.192
15	0.836	0.164	0.811	0.189	0.799	0.201
25	0.838	0.162	0.837	0.163	0.832	0.168
50	0.837	0.163	0.834	0.166	0.839	0.161
75	0.838	0.162	0.838	0.162	0.831	0.169
Mutual Words Removed	Teens	Adults	Teens	Adults	Teens	Adults
5	0.176	0.840	0.186	0.814	0.090	0.910
15	0.172	0.833	0.170	0.830	0.147	0.853
25	0.166	0.837	0.179	0.821	0.151	0.849
50	0.165	0.837	0.170	0.830	0.173	0.827
75	0.163	0.838	0.173	0.827	0.178	0.822

Table 35. Average Percentage of Use of the Mutual High-Frequency Stop N-grams.

In the teens versus 20s classification task, teens comprised of 40% of the authors (training and test sets), so there was an almost balanced data set. The usage of mutual high-frequency stop n-grams, however, was not balanced. In all the trigram experiments, teens used the stop n-grams less than 27% of the time.

Thus, the removal of n-grams that were used less often by teens could be the cause for the significant loss of performance. Even the removal of only five mutual n-grams caused the F-score to go from 0.744 to 0.581.

In both the teens versus 30s and teens versus 40s classification tasks, there was a more even usage of the stop n-grams, especially in the teens versus 40s classification task. There, in a majority of instances, the removal of the stop n-grams increased performance from the base feature where no stop n-grams were removed. In fact, the best F-scores are from features sets that remove mutual high-frequency stop n-grams.

Even though there was disproportionate usage of stop n-grams between teens and 50-year-olds, the performance of the SVM was still very good. A reason for this could be the differences in vocabulary between the two age groups permeated throughout the entire vocabulary. The NBC experiments support this reason. In the teens versus 50s experiments with the NBC using Witten-Bell smoothing, though the prior probability of the teen class is 85%, there was still an average precision of over 90%. The vocabulary of the 50-year-olds was different enough that it is able to overcome the high prior probability that an n-gram was written by a teen.

The usage of the mutual stop n-grams was uneven between teens and adults as well. There, however, was not a precipitous drop in performance as n-grams were removed. The removal of some mutual n-grams gave a slight edge in performance over not removing any n-grams at all. A possible reason for this may be because teens have a different enough vocabulary that the removal of some n-grams does not remove all the discriminating n-grams. As more and more n-grams were removed, however, there was a decrease in performance, especially in the trigram feature set. In the top 10 results, the unigram and bigram feature sets allowed for the removal of only up to 15 mutual n-grams.

It appears that the more even the usage of the stop n-grams, the better the performance by removing such n-grams; the more disparate the usage, the

less potential for a positive effect. Where there is a disparate usage of common n-grams, it may be better to use the entropy-generated words than high-frequency words. If the n-gram usage is almost even, it appears to be beneficial to remove those common n-grams. Future experiments are needed to explore the effects of removal of n-grams where there is an uneven usage, because in some cases, it is beneficial up to a certain point. Those experiments can also help determine the number of n-grams to remove before there is a significant decrease in performance.

5. Character N-grams

Character n-grams did not perform as well as word n-grams in the SVM model and only gave a slight edge to word n-grams in the NBC model. They did perform well in the NBC model in the teens versus 30s/40s/50s classification task. In some cases, different sized grams did slightly better. The character grams, however, did not do well when distinguishing between teens and adults. In this case, it could be that context with word phrases is more important. The success of trigrams when distinguishing adults and 20-year-olds from teens demonstrates the need for context.

In this research, only three *character* n-grams were used in the SVM experiments. Future experiments with different sized character n-grams could explore the effectives of this feature type to distinguish chat.

6. Meta-Data Features

The addition of the meta-data features in all cases degraded performance. The reason for this might be due to the small file sizes of some of the documents. In our corpus, there were 1,717 documents that were 1 kilobyte (kb) or less in size. Because of the small sizes, there may not be enough data to capture meaningful counts of the style-based features. An experiment removing all files less than 1 kb in size, was performed. The removal of such files, however, caused severe degradation to performance, due to the smaller training set.

In the research by Kucukyilmaz et al., style-based features, however, were the best feature type. Because they did not use a fine grain approach to counting the frequency of the style-based terms, they may have been able to get more meaningful feature values. Future experiments with similar categories, defined by thresholds that Kucukyilmaz et al. used as meta-data feature values, could improve performance [11]. Such experiments could also include other types of meta-data, such as the frequency of stop n-grams or misspelled words.

7. Lin Features

Using the Lin features set, the SVM model performed marginally better than Lin's NBC, when classifying teens versus 20-year-olds. The SVM model, however, did not generate better results, compared to Lin's NBC results for the other age groups. Table 36 contains Lin's best NBC results and the SVM results from this research.

Classification Task	NBC F-score	SVM F-score
Teens vs. 20s	0.464	0.468
Teens vs. 30s	0.786	0.785
Teens vs. 40s	0.814	0.827
Teens vs. 50s	0.932	0.922

Table 36. Comparison of SVM and NBC Models Using Lin's Feature Set [11].

Since Lin's features used a fine grain count of punctuation and emoticons—in essence a subset of meta-data features—using a less granular approach to represent the frequency of punctuation and emoticons could improve performance. Future experiments with a similar approach to Kucukyilmaz et al., using threshold values rather than frequency counts [11], may improve performance using this feature set.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS

A. SUMMARY

Both models demonstrate that they have the capability of distinguishing age groups. The SVM model, however, outperforms the NBC, because it is better able to handle the unbalanced data in the teens/adult data set. It is possible that with a more balanced teen/adult data set, the NBC's performance could improve. In almost all classifications tasks for both models, trigrams are the best feature type for both models.

The removal of stop n-grams sometimes improved performance, but at times also degraded performance. N-grams with entropy of 1.000 generated a more stable list of stop words than high-frequency-based n-grams. In most cases, the removal of such n-grams did not significantly affect performance. Because of the lack of effect on performance, this method could be used in future experiments to help decrease dimensionality in the feature vectors without concern for performance loss. Caution should be exercised, because in the few experiments where stop n-grams with entropy of less than 1.000 were removed, there was a noticeable decrease in performance.

Similarly, it appeared that the removal of high-frequency-based n-grams was beneficial, when the distribution of the stop n-gram across classes was uniform. If there is a disproportionate use, however, there is a risk of removing n-grams that are distinct and contain more information. An exploration of the effects of removing words at different thresholds is needed. If more words can be removed, it could improve the NBC, because words that have more distinctiveness would be given more weight. It may also improve the SVM model by reducing the amount of sparse data.

Inclusion of meta-data, or only using meta-data as a feature, produced poor results. The addition of such features may have caused an increase in sparse data to the feature vectors to the point of degrading performance. Instead of using a fine grain measurement of meta-data, categories defined by thresholds may generate better results.

B. FUTURE WORK

1. Exploration of Other Features/Kernels

None of the work in this study used combinations of feature types. The trigram feature type produced the best performance when classifying teens versus adults in the NBC model; but unigrams and bigrams produced the best results in the SVM model. A combination of using both n-gram types may result in an increase in performance. Also, future experiments can help determine the optimal threshold for stop n-gram removal of both entropy-based and high-frequency stop n-grams. Such experiments can also explore normalizing the entropy value and frequency counts to generate better stop n-gram lists. Though not found in this research, past research has shown that emoticons and punctuation do play a role in classifying age groups [11]. By using categories defined by thresholds, instead of individual frequency counts as meta-data feature values, the inclusion of meta-data could help fine-tune the classifiers for better performance.

Also, this research only used a linear kernel for the SVM. Given that conversations among teens and people in their early 20s are very similar, a linear kernel may not be able to produce a cleanly separating hyperplane. Other kernel types may be able to generate a better separation between the teens and 20s age groups, thus better able to classify them.

2. Deception of Age

The corpus used in our experiments relied on self-reported ages, which we assumed were true. When creating an online profile, people can easily

misrepresent their age. The models we generated have the potential to distinguish different age groups, but our experiments did not try to distinguish ages when a person pretended to be of a different age. Our models have success with truthful ages, but experiments with misrepresentative ages are needed to determine if SVMs and NBCs have the ability to distinguish the actual age.

3. Multi-class Classifier

A multi-class classifier would be able to handle a more realistic scenario of a general chat room, where people of all ages converse. This classifier could also help narrow observation of chat conversations to specific age groups. We attempted to perform multi-class classification using a linear kernel, but the results are not at all noteworthy. Because there is some similarity between age groups, they may occupy the same vector space areas, therefore not allowing for a clean linear division. This similarity is demonstrated by the small size of the slack variables used (e.g., in the case of teens versus 40s, the slack variable is 2^{-15}). SVM models using non-linear kernels may generate better performance.

4. Cross Domain into Instant Messaging

The YISS-2 survey found that 40% of the first incident of a sexual solicitation occurred when using Instant Messaging [2]. While similar to online chat, Instant Messaging is more akin to a private chat room channel for usually two, but sometimes more users. A method to detect adults conversing with teens is needed to prevent solicitations via Instant Messaging. Future experiments can explore the ability or fine-tune the models generated in this research to cross domain into Instant Messaging.

5. Detection of Distribution of Child Pornography

The YISS-2 survey also detected a new trend of solicitations, where 27% of solicitors asked youths for sexual photographs of themselves [2]. The National

Juvenile Online Victimization Study found that 40% of criminals arrested for possession of child pornography also sexually victimized children [25]. A reason why child pornography possessors may collect it is so they can use it to "groom children and lower their inhibitions [26]." When trying to detect suspicious behavior, the solicitation for sexual photographs increases the probability of a follow-on aggressive solicitation. Future research in detection of online predators should not only include detection of adults conversing with teens, but also include the detection of exchanges of sexual photos.

C. CONCLUDING REMARKS

The results of this research produced results that show that it is possible to differentiate between adult conversations and teen conversations, using models generated by a Support Vector Machine. More experiments, using combined or different feature types; other machine learning techniques; and crossing different online media domains, are needed to further improve performance to the point where an automatic detection system can be fielded. Given the lack of reporting to parents and law enforcement, there is a great need for an automated system to help prevent online solicitations of youths from becoming offline victims.

APPENDIX A: SUPPORT VECTOR MACHINE

The Support Vector Machine is discussed in [10, 18–21]. Also known as a maximum margin classifier, the SVM tries to find the line between two classes of data that maximizes the margin between them. Because data being classified is not always linearly separable (i.e., there is not always a line, or hyperplane, which can separate the two classes of data), it is often transformed using a kernel function. Though there are different types of kernels for an SVM, this research used a linear kernel, which does not transform the data. The two classes of data are represented by n-dimensional vectors, where each dimension represents a feature, such as an n-gram. Using the training set vectors, the SVM generates the hyperplane (model vector) that separates the two classes with the maximum margin. The test set vectors and model vector are then used to determine which side of the hyperplane the test vectors lay. The side that a test vector lies upon is its predicted class.

Based on the training data, a SVM will find the maximum margin hyperplane that separates the two classes. A maximum margin hyperplane exists where the distance from the closest data point to the hyperplane is as large as possible. Support vectors are the data points that are on the margin. Figure 6 is an example of a hyperplane that creates the maximum margin between classes. Also, in Figure 6, the support vectors are circled. The maximum margin in the figure is the distance between lines l_1 and l_2.

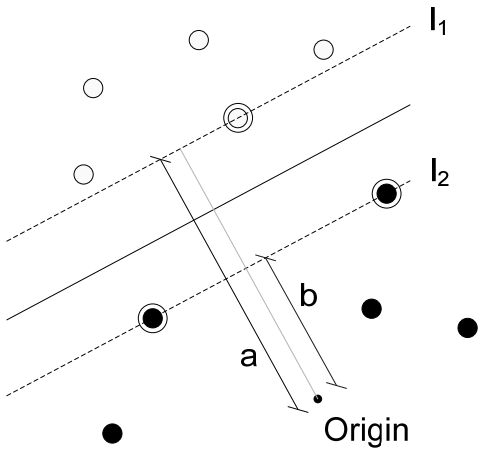

Figure 6. Linear Separating Hyperplanes.

A. DETERMINING THE SIZE OF THE MARGIN

The first step in defining the maximum margin hyperplane, is to find the size of the margin. The equation for a hyperplane is $\vec{w}^T\vec{x}+b=0$. If the hyperplane is to separate data, then its equation will be $\vec{w}^T\vec{x}_k+b>0$ for all $\vec{x}_k$ of one class and $\vec{w}^T\vec{x}_j+b<0$ for all $\vec{x}_j$ of the other class [21]. Let the training points be labeled as $y_k\in\{-1,1\}$, with 1 being a positive example and -1 being a negative example, thus the hyperplane can be defined as

$y_k(\vec{w}^T\vec{x}_k+b)\geq 0$ for all points.

Because $\frac{b}{\|\vec{w}\|}$ determines the hyperplane's offset from the origin along the vector $\vec{w}$, $\vec{w}$ and b can be scaled without changing the hyperplane. To prevent such scaling, $\vec{w}$ and b are chosen such that

$y_k(\vec{w}^T\vec{x}_k+b)\geq 1 \ \ \forall k$.

To help find the separating hyperplane, think of two hyperplanes, represented by l_1 and l_2 in Figure 6, which are parallel to the separating hyperplane. The points of one class closest to points of the other class that lie on these planes are also known as support vectors. Because of our choice of

values for $\vec{w}$ and b, the equations for these planes are $y_k(\vec{w}^T\vec{x}_k + b) = 1$ and $y_j(\vec{w}^T\vec{x}_j + b) = 1$ for some points j, k where j is a data point for the positive class ($y_j = 1$) and k is a data point for the negative class ($y_k = -1$). There could be more than one point lying on these planes. The separating hyperplane's distance to the margin is then half the distance between l_1 and l_2.

This distance between l_1 and l_2 is also the same as the difference in distance from the origin to the closet point to l_1 and the distance from the origin to the closet point to l_2. The distance from the origin to the closest point on a hyperplane is found by minimizing $\vec{x}^T\vec{x}$ subject to $\vec{x}$ being on the hyperplane [21],

$$\min_{\|\vec{x}\|} \vec{x}^T\vec{x} + \lambda(\vec{w}^T\vec{x} + b - 1)$$

$$\frac{d}{d\vec{x}} = 0 = 2\vec{x} + \lambda\vec{w} = 0 \qquad \text{therefore,}$$

$$\vec{x} = -\frac{\lambda}{2}\vec{w} \qquad \text{substituting this } \vec{x} \text{ into } w^T\vec{x} + b - 1 = 0 \text{ results in}$$

$$-\frac{\lambda}{2}\vec{w}^T\vec{w} + b = 1 \qquad \text{therefore,}$$

$$\lambda = \frac{2(b-1)}{\vec{w}^T\vec{w}} \qquad \text{substituting this } \lambda \text{ into } \vec{x} = -\frac{\lambda}{2}\vec{w} \text{ results in}$$

$$\vec{x} = \frac{1-b}{\vec{w}^T\vec{w}}\vec{w}$$

$$\vec{x}^T\vec{x} = \frac{(1-b)^2}{(\vec{w}^T\vec{w})^2}\vec{w}^T\vec{w} = \frac{(1-b)^2}{\vec{w}^T\vec{w}}$$

$$\|\vec{x}\| = \sqrt{\vec{x}^T\vec{x}} = \frac{|1-b|}{\sqrt{\vec{w}^T\vec{w}}} = \frac{|1-b|}{\|\vec{w}\|}$$

$$\|\vec{x}\| = \frac{|-1-b|}{\|\vec{w}\|} \qquad \text{result of working out } \vec{w}^T\vec{x} + b = -1.$$

Subtracting these two distances gives the margin size,

$$\frac{|1-b|}{\|\vec{w}\|} - \frac{|-1-b|}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|}.$$

B. DEFINING THE MAXIMUM MARGIN HYPERPLANE

To maximize the size of the margin, $\frac{2}{\|\vec{w}\|}$, which would give the greatest distance between the classes, the denominator, $\|\vec{w}\|$, must be minimized, subject to the constraint that $y_k(\vec{w}^T\vec{x}_k + b) \geq 1 \ \ \forall k$.

To do so, we use the Karush Kuhn Tucker (KKT) setup using positive Lagrange multipliers and subtract the constraints. Because both the main term and the constraints are linear convex, this becomes a convex quadratic optimization problem [10] where

$$\begin{aligned} L_p &= \frac{1}{2}\vec{w}^T\vec{w} - \sum_k \lambda_k (y_k(\vec{w}^T\vec{x}_k + b) - 1) \\ &= \frac{1}{2}\vec{w}^T\vec{w} - \sum_k \lambda_k y_k(\vec{w}^T\vec{x}_k + b) + \sum_k \lambda_k \\ &= \frac{1}{2}\vec{w}^T\vec{w} - \sum_k \lambda_k y_k \vec{w}^T\vec{x}_k - \sum_k \lambda_k y_k b + \sum_k \lambda_k \\ &= \frac{1}{2}\vec{w}^T\vec{w} - \vec{w}^T\sum_k \lambda_k y_k \vec{x}_k - b\sum_k \lambda_k y_k + \sum_k \lambda_k . \end{aligned}$$

Using the KKT conditions, we can then solve the dual problem, which is to maximize L_p with respect to λ_k, subject to the constraints that the gradient of L_p with respect to $\vec{w}$ and b are 0 and that $\lambda_k \geq 0$.

$$\frac{\delta L_p}{\delta \vec{w}} = 0 \Rightarrow \vec{w} = \sum_k \lambda_k y_k \vec{x}_k$$

$$\frac{\delta L_p}{\delta b} = 0 \Rightarrow \sum_k \lambda_k y_k$$

Substituting the above into L_p, we get the dual

$$L_d = \frac{1}{2}(\vec{w}^T \vec{w}) - \vec{w}^T \sum_k \lambda_k y_k \vec{x}_k - b\sum_k \lambda_k y_k + \sum_k \lambda_k$$
$$= -\frac{1}{2}(\vec{w}^T \vec{w}) + \sum_k \lambda_k$$
$$= -\frac{1}{2}\sum_k \sum_l \lambda_k \lambda_l y_k y_l (\vec{x}_k)^T \vec{x}_l + \sum_k \lambda_k$$

which is maximized with respect to λ_k, subject to the constraints $\sum_k \lambda_k y_k = 0$, and $\lambda_k \geq 0, \ \forall k$. Quadratic optimization methods are used to solve the above equation [10]. Once one solves for λ_k, the $\lambda_k > 0$ are the support vectors and lie on the separating hyperplanes. All other training points have $\lambda_k = 0$ and lie either on the separating hyperplanes, or in the classification region. The support vectors satisfy the equation $y_k(\vec{w}^T \vec{x}_k + b) = 1$. Thus b is solved for by finding one of the active constraints $y_k(\vec{w}^T \vec{x}_k + b) \geq 1$ where λ_k is not zero [21]. To maintain numerical stability, b is calculated for every support vector and the average value is used [10]. Once $\vec{w}$ and b are known, the separating hyperplane is found.

It is unlikely that the hyperplane will cleanly separate the data in real world problems. Classes are likely to overlap or have a very small margin. "Slack variables" compensate for this effect [21]. Rather than having $y_k(\vec{w}^T \vec{x}_k + b) \geq 1 \ \ \forall k$, slack variables, s_k, are introduce such that

$$y_k(\vec{w}^T \vec{x}_k + b) \geq 1 + s_k \ \ \forall k .$$

Without violating the constraint above, the slack variable allows a point to be s_k distance on the wrong side of the hyperplane. To prevent large slack variables from allowing any line to partition the data, another term is added to the Lagrangian to penalize large slacks [21],

$$L_p = \frac{1}{2}\vec{w}^T \vec{w} - \sum \lambda_k (y_k(\vec{w}^T \vec{x}_k + b) + s_k - 1) + \alpha \sum s_k .$$

The equation is then minimized as above.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B: ENTROPY-BASED STOP WORD LISTS

This appendix contains the 75 n-grams in the entropy-based stop n-gram lists for the seventh random training set. This set contained n-grams that had entropy values of less than 1.000, which may have caused degradation to the SVM's classification performance. Tables 37–51 contain the entropy value as well as the adult and teen usage per n-gram. The n-grams are ordered by their entropy, in ascending order.

Teens Versus 20s			
Unigram	Entropy	20s	Teen
hoho.	1.000	2	2
rarely	1.000	2	2
;-)	1.000	88	88
nooooo	1.000	3	3
ewwwwwww	1.000	2	2
funny..	1.000	2	2
rate	1.000	2	2
papi.	1.000	6	6
sting	1.000	2	2
ghost	1.000	4	4
2endsoftime	1.000	5	5
tail?	1.000	2	2
object	1.000	2	2
character	1.000	4	4
foreign	1.000	2	2
skate	1.000	3	3
fried	1.000	4	4
knife	1.000	3	3
dude..	1.000	2	2
immature	1.000	4	4
booze	1.000	3	3
to..	1.000	2	2
experience	1.000	2	2
nuts.	1.000	2	2
takin	1.000	3	3
wall.	1.000	3	3
t.v.	1.000	2	2
ya!	1.000	3	3
cheek	1.000	3	3
shopping	1.000	4	4
messing	1.000	3	3
md	1.000	4	4
whores	1.000	3	3
shudders.	1.000	2	2
mom,	1.000	3	3
trail	1.000	3	3
pa?	1.000	3	3
pisses	1.000	3	3

Teens Versus 20s			
Unigram	Entropy	20s	Teen
brat	1.000	3	3
cereal	1.000	4	4
bangs	1.000	2	2
nods.	1.000	4	4
theory	1.000	5	5
friends,	1.000	2	2
*throws	1.000	2	2
funniest	1.000	2	2
redorchid	1.000	5	5
stoned	1.000	3	3
sneezes.	1.000	2	2
wide	1.000	3	3
fields	1.000	2	2
not?	1.000	5	5
sn	1.000	2	2
kansas	1.000	2	2
gay!	1.000	3	3
it"	1.000	2	2
phat?	1.000	3	3
wait,"	1.000	5	5
loads	1.000	2	2
else?	1.000	2	2
jus	1.000	14	14
porkpiehat	1.000	2	2
teen	1.000	2	2
dumb.	1.000	2	2
addict	1.000	2	2
career	1.000	2	2
phat,"	1.000	2	2
creative	1.000	2	2
gummy	1.000	2	2
skool	1.000	2	2
males	1.000	5	5
what?!	1.000	2	2
think.	1.000	6	6
sucked	1.000	6	6
ther	1.000	4	4

Table 37. Entropy-Based Stop Unigrams for Teens vs. 20s Classification Task.

Teens Versus 30s			
Unigram	Entropy	30s	Teen
pants.	1.000	4	4
load	1.000	2	2
mi	1.000	3	3
nuttin	1.000	3	3
song	1.000	2	2
been?	1.000	6	6
ewwwwwww	1.000	2	2
rate	1.000	2	2
guest	1.000	2	2
cheerleader	1.000	2	2
rose	1.000	4	4
party.	1.000	2	2
sting	1.000	2	2
attractive	1.000	2	2
personal	1.000	5	5
fat.	1.000	2	2
parties	1.000	3	3
tail?	1.000	2	2
thing.	1.000	3	3
will	1.000	2	2
sleep	1.000	2	2
they?	1.000	3	3
disease	1.000	2	2
room..	1.000	2	2
matching	1.000	3	3
chilly	1.000	4	4
cotton	1.000	2	2
conversations	1.000	2	2
nuts.	1.000	2	2
takin	1.000	3	3
harsh	1.000	4	4
ireland	1.000	3	3
bare	1.000	2	2
school	1.000	2	2
shopping	1.000	4	4
ridiculous	1.000	2	2
hot	1.000	2	2
mom	1.000	3	3

Teens Versus 30s			
Unigram	Entropy	30s	Teen
trail	1.000	3	3
msn?	1.000	2	2
um	1.000	2	2
affection	1.000	2	2
hya	1.000	2	2
rained	1.000	2	2
hides.	1.000	5	5
runs.	1.000	2	2
pushes	1.000	2	2
sicko	1.000	2	2
funniest	1.000	2	2
eggs	1.000	2	2
clown	1.000	3	3
sneezes.	1.000	2	2
spanks	1.000	2	2
awful	1.000	3	3
canada?	1.000	3	3
once	1.000	2	2
gah	1.000	2	2
it"	1.000	2	2
fuker	1.000	2	2
why's	1.000	2	2
seen	1.000	52	52
blankie	1.000	2	2
ooo	1.000	2	2
rode	1.000	3	3
scar	1.000	4	4
attitude	1.000	2	2
george	1.000	4	4
hitting	1.000	4	4
snl	1.000	2	2
baby!	1.000	2	2
said.	1.000	4	4
career	1.000	2	2
creative	1.000	2	2
pfffft	1.000	2	2
specific	1.000	2	2

Table 38. Entropy-Based Stop Unigrams for Teens vs. 30s Classification Task.

Teens Versus 40s			
Unigram	Entropy	40s	Teen
load	1.000	2	2
fair	1.000	4	4
breath	1.000	4	4
omg..	1.000	2	2
sting	1.000	2	2
yes!	1.000	4	4
whenever	1.000	2	2
pie.	1.000	2	2
too..	1.000	4	4
dodge	1.000	2	2
plain	1.000	2	2
no?	1.000	4	4
mothman	1.000	2	2
blast	1.000	2	2
him???	1.000	2	2
alrighty	1.000	2	2
hush	1.000	8	8
mom	1.000	3	3
harder	1.000	2	2
bangs	1.000	2	2
buys	1.000	2	2
personally	1.000	2	2
money	1.000	2	2
hmmmmmmmm	1.000	3	3
tank	1.000	2	2
spanks	1.000	2	2
jersey	1.000	4	4
penny	1.000	3	3
steal	1.000	2	2
sits	1.000	31	31
wind?	1.000	2	2
rub	1.000	2	2
joke	1.000	13	13
anti	1.000	2	2
leather	1.000	2	2
around.	1.000	3	3
gorilla	1.000	2	2
teen	1.000	2	2

Teens Versus 40s			
Unigram	Entropy	40s	Teen
sooo	1.000	13	13
ull	1.000	2	2
specific	1.000	2	2
worried	1.000	2	2
today...	1.000	2	2
choice	1.000	4	4
panties	1.000	2	2
baby	1.000	35	35
un	1.000	3	3
king	1.000	10	10
animals	1.000	4	4
coach	1.000	2	2
<~~	1.000	2	2
wipes	1.000	3	3
rush	1.000	3	3
dangerous	1.000	2	2
krista	1.000	2	2
tan	1.000	3	3
cell	1.000	2	2
boom	1.000	2	2
sentence	1.000	2	2
dancin	1.000	2	2
minutes	1.000	8	8
box	1.000	8	8
ks	1.000	2	2
u?	1.000	17	17
bet	1.000	18	18
that?	1.000	20	20
hawaii	1.000	3	3
chicks	1.000	6	6
note	1.000	3	3
western	1.000	2	2
asking	1.000	19	19
officially	1.000	3	3
automatically	1.000	2	2
yesss	1.000	2	2
nh	1.000	5	5

Table 39. Entropy-Based Stop Unigrams for Teens vs. 40s Classification Task.

Teens Versus 50s			
Unigram	Entropy	50s	Teen
ok..	1.000	3	3
sting	1.000	2	2
holding	1.000	2	2
monday	1.000	2	2
bottom	1.000	2	2
plain	1.000	2	2
veronica	1.000	3	3
scarlet	1.000	3	3
cook	1.000	4	4
now...	1.000	3	3
vegas	1.000	3	3
bob	1.000	8	8
marie	1.000	2	2
harder	1.000	2	2
ain't	1.000	2	2
rained	1.000	2	2
money	1.000	2	2
woman	1.000	10	10
stands	1.000	2	2
chains	1.000	2	2
kansas	1.000	2	2
absolutely	1.000	2	2
flying	1.000	2	2
hanging	1.000	2	2
bumps	1.000	2	2
go	1.000	2	2
barely	1.000	2	2
glove	1.000	2	2
tosses	1.000	3	3
panties	1.000	2	2
hobo	1.000	3	3
giggles.	1.000	2	2
sam	1.000	3	3
uses	1.000	2	2
know..	1.000	2	2
atm	1.000	5	5
waves	1.000	2	2
round	1.000	4	4

Teens Versus 50s			
Unigram	Entropy	50s	Teen
hahahahah	1.000	2	2
spent	1.000	2	2
raining	1.000	2	2
titty	1.000	2	2
gawd	1.000	2	2
<<	1.000	3	3
lived	1.000	6	6
farm	1.000	2	2
skank	1.000	2	2
????	1.000	2	2
room!	1.000	3	3
dione	1.000	3	3
due	1.000	2	2
changes	1.000	2	2
calif	1.000	4	4
too?	1.000	2	2
lives	1.000	5	5
les	1.000	2	2
moving	1.000	3	3
whispers	1.000	2	2
mouse	1.000	2	2
lap	1.000	4	4
pair	1.000	2	2
checked	1.000	2	2
0.999411065	0.999	17	18
hi	0.999	333	358
slow	0.998	11	10
peace	0.998	9	10
yo	0.995	21	25
cough	0.994	6	5
lizzie	0.994	5	6
snow	0.994	5	6
muh	0.994	5	6
rose	0.991	5	4
near	0.991	8	10
american	0.991	4	5
far	0.991	8	10

Table 40. Entropy-Based Stop Unigrams for Teens vs. 50s Classification Task.

Teens Versus Adults			
Unigram	Entropy	Adult	Teen
hoho.	1.000	2	2
funny..	1.000	2	2
drunk?	1.000	3	3
shhh	1.000	3	3
mark	1.000	4	4
late.	1.000	4	4
object	1.000	2	2
scott	1.000	3	3
scoot	1.000	2	2
ours	1.000	2	2
people!	1.000	3	3
you*	1.000	2	2
combination	1.000	2	2
models	1.000	3	3
ridiculous	1.000	2	2
jerks	1.000	2	2
3	1.000	2	2
21?	1.000	2	2
drawing	1.000	2	2
fields	1.000	2	2
sn	1.000	2	2
gah	1.000	2	2
thousands	1.000	2	2
scar	1.000	4	4
grades	1.000	2	2
pvt	1.000	2	2
teens	1.000	4	4
career	1.000	2	2
lol!!!	1.000	2	2
phat	1.000	2	2
oooh.	1.000	2	2
wires	1.000	2	2
helmet	1.000	3	3
heels	1.000	2	2
heritage	1.000	2	2
mt	1.000	2	2
bunny.	1.000	2	2
maineman1701	1.000	2	2

Teens Versus Adults			
Unigram	Entropy	Adult	Teen
switches	1.000	2	2
rally	1.000	2	2
hai	1.000	2	2
curled	1.000	2	2
itself	1.000	2	2
vans	1.000	3	3
cap	1.000	2	2
highschool	1.000	2	2
ud	1.000	2	2
czech	1.000	3	3
sales	1.000	2	2
!shot	1.000	2	2
gays	1.000	2	2
chews	1.000	2	2
thigh.	1.000	2	2
jade	1.000	2	2
gore	1.000	2	2
conversation	1.000	2	2
grins.	1.000	2	2
dro	1.000	2	2
daniels	1.000	3	3
fukin	1.000	2	2
gt	1.000	4	4
crawl	1.000	2	2
branch	1.000	2	2
attention.	1.000	2	2
nighters	1.000	3	3
laffs	1.000	2	2
niks	1.000	2	2
rifle	1.000	2	2
argue	1.000	8	8
keeley	1.000	2	2
now!!!	1.000	2	2
face..	1.000	2	2
pool.	1.000	2	2
daniel	1.000	4	4
na	1.000	3	3

Table 41. Entropy-Based Stop Unigrams for Teens vs. Adults Classification Task.

Teens Versus 20s			
Bigram	Entropy	20s	Teen
you believe	1.000	2	2
<post> oooh.	1.000	2	2
your asl?	1.000	2	2
give u	1.000	3	3
come and	1.000	3	3
most women	1.000	3	3
because my	1.000	3	3
dont you	1.000	6	6
moans. </post>	1.000	3	3
puts his	1.000	2	2
<post> pfft!	1.000	2	2
pa? </post>	1.000	3	3
dogs </post>	1.000	2	2
hair is	1.000	3	3
all girls	1.000	2	2
subject </post>	1.000	2	2
whoa </post>	1.000	7	7
sure is	1.000	2	2
miss my	1.000	2	2
or they	1.000	2	2
back.. </post>	1.000	2	2
tail? </post>	1.000	2	2
glad i	1.000	4	4
could </post>	1.000	3	3
<post> 10	1.000	2	2
not there	1.000	2	2
army </post>	1.000	2	2
white people	1.000	4	4
any1 from	1.000	2	2
peaches for	1.000	2	2
up her	1.000	2	2
cat. </post>	1.000	3	3
cant remember	1.000	2	2
<post> talk	1.000	8	8
stupid. </post>	1.000	3	3
wall. </post>	1.000	3	3
monster </post>	1.000	2	2
make us	1.000	2	2

Teens Versus 20s			
Bigram	Entropy	20s	Teen
i may	1.000	4	4
sumthin </post>	1.000	2	2
bday is	1.000	2	2
sure? </post>	1.000	2	2
size of	1.000	2	2
op. </post>	1.000	2	2
brown </post>	1.000	2	2
.action plays	1.000	6	6
nuts. </post>	1.000	2	2
hate when	1.000	6	6
<post> pm	1.000	37	37
is older	1.000	2	2
to roll	1.000	2	2
beast </post>	1.000	2	2
a waste	1.000	4	4
the inside	1.000	3	3
who me?	1.000	3	3
i spent	1.000	2	2
runs around	1.000	3	3
pay for	1.000	2	2
asl? </post>	1.000	9	9
appreciate it	1.000	2	2
for life	1.000	2	2
smoke. </post>	1.000	3	3
girlfriend </post>	1.000	5	5
out if	1.000	2	2
the pool	1.000	2	2
thought i'd	1.000	2	2
sauce </post>	1.000	2	2
would just	1.000	4	4
keep my	1.000	2	2
paying attention.	1.000	2	2
my b-day	1.000	2	2
gonna pm	1.000	3	3
the face.	1.000	3	3
chocolate </post>	1.000	2	2
cause if	1.000	2	2

Table 42. Entropy-Based Stop Bigrams for Teens vs. 20s Classification Task.

Teens Versus 30s			
Bigram	Entropy	30s	Teen
people like	1.000	2	2
dirty </post>	1.000	3	3
me. i	1.000	2	2
girl! </post>	1.000	3	3
one who	1.000	2	2
u ever	1.000	2	2
kid is	1.000	2	2
ran out	1.000	2	2
puts his	1.000	2	2
dogs </post>	1.000	2	2
can call	1.000	5	5
i dunno,	1.000	3	3
sure is	1.000	2	2
miss my	1.000	2	2
tail? </post>	1.000	2	2
on what	1.000	4	4
glad i	1.000	4	4
of going	1.000	2	2
the eyes	1.000	2	2
mrqd </post>	1.000	2	2
mean he	1.000	2	2
tall </post>	1.000	2	2
lol me	1.000	2	2
guy </post>	1.000	19	19
could i	1.000	2	2
dont mind	1.000	5	5
hit my	1.000	2	2
for some	1.000	11	11
heck is	1.000	2	2
i do.	1.000	2	2
<post> 24	1.000	2	2
.action beats	1.000	4	4
any nice	1.000	3	3
clue </post>	1.000	2	2
back is	1.000	2	2
your life	1.000	2	2
any way	1.000	2	2
my hands	1.000	2	2

Teens Versus 30s			
Bigram	Entropy	30s	Teen
oops sorry	1.000	2	2
brown </post>	1.000	2	2
ridiculous </post>	1.000	2	2
then get	1.000	2	2
and youre	1.000	3	3
on all	1.000	2	2
that there	1.000	2	2
<post> keep	1.000	3	3
will i	1.000	2	2
like any	1.000	2	2
guy was	1.000	3	3
the inside	1.000	3	3
to kill	1.000	4	4
cant take	1.000	2	2
didnt know	1.000	11	11
the face	1.000	4	4
dances around	1.000	2	2
do it?	1.000	2	2
boys </post>	1.000	3	3
out if	1.000	2	2
dammit i	1.000	2	2
thought i'd	1.000	2	2
.action steals	1.000	2	2
myself </post>	1.000	11	11
scott </post>	1.000	14	14
and i've	1.000	3	3
keep my	1.000	2	2
air. </post>	1.000	2	2
like i'm	1.000	2	2
i are	1.000	3	3
take you	1.000	3	3
<post> gets	1.000	2	2
her a	1.000	4	4
<post> bless	1.000	2	2
it good	1.000	2	2
the bunny	1.000	3	3
he cant	1.000	2	2

Table 43. Entropy-Based Stop Bigrams for Teens vs. 30s Classification Task.

Teens Versus 40s			
Bigram	Entropy	40s	Teen
in about	1.000	2	2
<post> er	1.000	2	2
why would	1.000	4	4
mine is	1.000	3	3
one who	1.000	2	2
girl lol	1.000	2	2
<post> nope,	1.000	2	2
<post> bbl	1.000	9	9
i can	1.000	65	65
dogs </post>	1.000	2	2
is love	1.000	2	2
she wants	1.000	3	3
miss my	1.000	2	2
glad i	1.000	4	4
<post> 10	1.000	2	2
where do	1.000	3	3
mrqd </post>	1.000	2	2
not there	1.000	2	2
state </post>	1.000	3	3
awake </post>	1.000	2	2
it, i	1.000	2	2
ago </post>	1.000	11	11
hit my	1.000	2	2
may not	1.000	3	3
<post> didn't	1.000	4	4
size of	1.000	2	2
mate </post>	1.000	3	3
why they	1.000	2	2
of all	1.000	2	2
convo </post>	1.000	3	3
.action cries.	1.000	3	3
oops sorry	1.000	2	2
worry about	1.000	2	2
florida </post>	1.000	7	7
tried to	1.000	3	3
right there	1.000	2	2
him??? </post>	1.000	2	2
.action cries	1.000	2	2

Teens Versus 40s			
Bigram	Entropy	40s	Teen
car </post>	1.000	3	3
i ain't	1.000	2	2
<post> shit	1.000	3	3
was getting	1.000	2	2
he went	1.000	2	2
do it?	1.000	2	2
thought i'd	1.000	2	2
eh? </post>	1.000	7	7
albany </post>	1.000	2	2
myself </post>	1.000	11	11
the place	1.000	2	2
thinking about	1.000	4	4
in it	1.000	12	12
wb gaston	1.000	2	2
came to	1.000	4	4
an ass	1.000	2	2
say anything	1.000	2	2
u get	1.000	7	7
end of	1.000	4	4
he cant	1.000	2	2
a baby	1.000	2	2
afk for	1.000	2	2
with it	1.000	6	6
me when	1.000	4	4
from new	1.000	2	2
let her	1.000	3	3
emma </post>	1.000	4	4
my pm	1.000	7	7
<post> nice,	1.000	3	3
called you	1.000	2	2
up like	1.000	2	2
time? </post>	1.000	2	2
said she	1.000	2	2
and good	1.000	3	3
she said	1.000	5	5
idea </post>	1.000	3	3
went out	1.000	2	2

Table 44. Entropy-Based Stop Bigrams for Teens vs. 40s Classification Task.

Teens Versus 50s			
Bigram	Entropy	50s	Teen
people who	1.000	2	2
you be	1.000	2	2
<post> hahahahah	1.000	2	2
than a	1.000	2	2
think the	1.000	2	2
is love	1.000	2	2
same here	1.000	2	2
the person	1.000	2	2
strat </post>	1.000	4	4
monster </post>	1.000	2	2
week </post>	1.000	2	2
<post> hahahah	1.000	3	3
is more	1.000	2	2
contented </post>	1.000	4	4
<post> getting	1.000	2	2
right there	1.000	2	2
own </post>	1.000	3	3
couple of	1.000	2	2
where it	1.000	3	3
yes or	1.000	2	2
<post> sometimes	1.000	2	2
in portland	1.000	2	2
got ya	1.000	2	2
.action giggles.	1.000	2	2
hello to	1.000	2	2
in and	1.000	2	2
back in	1.000	5	5
said she	1.000	2	2
the bottom	1.000	2	2
tired of	1.000	2	2
giggles. </post>	1.000	2	2
<post> ok..	1.000	3	3
was there	1.000	2	2
life is	1.000	2	2
sounds like	1.000	6	6
<post> lo	1.000	2	2
sam </post>	1.000	2	2
they got	1.000	2	2

Teens Versus 50s			
Bigram	Entropy	50s	Teen
would like	1.000	2	2
near the	1.000	2	2
snow </post>	1.000	3	3
the fun	1.000	2	2
canada </post>	1.000	2	2
my head	1.000	3	3
???? </post>	1.000	2	2
as the	1.000	2	2
:-) :-)	1.000	10	10
watch your	1.000	2	2
<post> papi	1.000	2	2
dione </post>	1.000	3	3
yikes </post>	1.000	2	2
<post> loves	1.000	2	2
in 3	1.000	2	2
cards </post>	1.000	2	2
am back	1.000	3	3
matt </post>	1.000	2	2
now you	1.000	2	2
so true	1.000	2	2
use to	1.000	3	3
thought u	1.000	2	2
lil </post>	1.000	4	4
tell them	1.000	2	2
hope not	1.000	2	2
remember when	1.000	2	2
has no	1.000	2	2
one day	1.000	2	2
when is	1.000	3	3
and talk	1.000	2	2
you shouldn't	1.000	2	2
when he	1.000	3	3
since when	1.000	3	3
a white	1.000	2	2
<post> <	1.000	2	2
me either	1.000	3	3
who can	1.000	2	2

Table 45. Entropy-Based Stop Bigrams for Teens vs. 50s Classification Task.

Teens Versus Adults			
Bigram	Entropy	Adult	Teen
dont ya	1.000	2	2
nooo </post>	1.000	8	8
cyber? </post>	1.000	3	3
most women	1.000	3	3
moans. </post>	1.000	3	3
sorry! </post>	1.000	3	3
all girls	1.000	2	2
whoa </post>	1.000	7	7
<post> nah.	1.000	3	3
kick me	1.000	2	2
ur name	1.000	5	5
any1 from	1.000	2	2
peaches for	1.000	2	2
sumthin </post>	1.000	2	2
op. </post>	1.000	2	2
ridiculous </post>	1.000	2	2
out! </post>	1.000	4	4
i seem	1.000	3	3
him??? </post>	1.000	2	2
beast </post>	1.000	2	2
cries.. </post>	1.000	2	2
runs around	1.000	3	3
sauce </post>	1.000	2	2
hiya starr	1.000	2	2
wb gaston	1.000	2	2
the face.	1.000	3	3
it worth	1.000	2	2
get gagged	1.000	2	2
is sick	1.000	3	3
like not	1.000	2	2
dumb as	1.000	2	2
<post> wait,	1.000	3	3
u hear	1.000	2	2
so bored.	1.000	2	2
its okay	1.000	2	2
<post> what??	1.000	3	3
.action moans.	1.000	3	3
me tonight	1.000	2	2

Teens Versus Adults			
Bigram	Entropy	Adult	Teen
shuts up.	1.000	2	2
to lie	1.000	2	2
balls. </post>	1.000	4	4
one up	1.000	2	2
yellow </post>	1.000	2	2
and ill	1.000	2	2
and even	1.000	2	2
to only	1.000	2	2
blanket </post>	1.000	2	2
beleive in	1.000	2	2
kid named	1.000	2	2
shot in	1.000	2	2
else who	1.000	2	2
be sure	1.000	3	3
hello hello	1.000	3	3
coming from	1.000	2	2
ignored </post>	1.000	2	2
where have	1.000	4	4
but really	1.000	2	2
<post> soooo	1.000	3	3
<post> what'd	1.000	2	2
how's everyone	1.000	2	2
of music	1.000	3	3
michael? </post>	1.000	3	3
crap outta	1.000	2	2
to mike	1.000	2	2
<post> ooo,	1.000	2	2
fresh </post>	1.000	2	2
german </post>	1.000	3	3
they both	1.000	2	2
like he	1.000	3	3
bit with	1.000	3	3
parents were	1.000	2	2
would if	1.000	2	2
song about	1.000	2	2
fans in	1.000	2	2
who hates	1.000	2	2

Table 46. Entropy-Based Stop Bigrams for Teens vs. Adults Classification Task.

Teens Versus 20s			
Trigram	Entropy	20s	Teen
u want me	1.000	3	3
<post> umm </post>	1.000	3	3
i feel a	1.000	2	2
talked to you	1.000	2	2
<post> but hey	1.000	3	3
<post> yeah lol	1.000	2	2
just got a	1.000	2	2
<post> mhmm </post>	1.000	3	3
talk to </post>	1.000	4	4
<post> u wish	1.000	2	2
in here lol	1.000	2	2
<post> u said	1.000	2	2
just like the	1.000	2	2
<post> is he	1.000	3	3
<post> my name	1.000	5	5
he called me	1.000	2	2
<post> .action sneezes.	1.000	2	2
<post> .action cant	1.000	2	2
give it up	1.000	4	4
<post> too </post>	1.000	2	2
<post> my profile	1.000	2	2
sup peeps </post>	1.000	2	2
<post> hey!! </post>	1.000	2	2
in here i	1.000	2	2
<post> lol heaven	1.000	3	3
doing it </post>	1.000	2	2
<post> .action doesn't	1.000	4	4
<post> because my	1.000	2	2
<post> need a	1.000	2	2
in it for	1.000	2	2
<post> i speak	1.000	2	2
i swear to	1.000	4	4
pretty cool </post>	1.000	2	2
would you like	1.000	2	2
4 years </post>	1.000	2	2
<post> .action sneezes	1.000	2	2
is why i	1.000	3	3
remember me </post>	1.000	2	2

Teens Versus 20s			
Trigram	Entropy	20s	Teen
and the other	1.000	3	3
i was younger	1.000	2	2
i cant do	1.000	3	3
the house </post>	1.000	3	3
that was the	1.000	2	2
a bunch of	1.000	6	6
to take my	1.000	2	2
didn't want to	1.000	2	2
<post> ass </post>	1.000	2	2
the size of	1.000	2	2
<post> .action got	1.000	2	2
hey mike </post>	1.000	2	2
girls wana chat	1.000	3	3
then go to	1.000	2	2
so i have	1.000	2	2
<post> heyy </post>	1.000	2	2
for a little	1.000	3	3
<post> hmmm... </post>	1.000	2	2
<post> .action wants	1.000	11	11
was just about	1.000	2	2
guys and girls	1.000	2	2
he has to	1.000	2	2
i pm you	1.000	3	3
any girl from	1.000	2	2
<post> im great	1.000	2	2
still here </post>	1.000	3	3
thought i was	1.000	5	5
he had a	1.000	3	3
where u from	1.000	3	3
a shower </post>	1.000	2	2
<post> .action moans.	1.000	3	3
<post> m </post>	1.000	2	2
<post> hey brb	1.000	2	2
paying attention. </post>	1.000	2	2
<post> hell yeah	1.000	5	5
<post> hey does	1.000	2	2
gave me a	1.000	2	2

Table 47. Entropy-Based Stop Trigrams for Teens vs. 20s Classification Task.

Teens Versus 30s			
Trigram	Entropy	30s	Teen
is just a	1.000	3	3
u lol </post>	1.000	2	2
<post> what kind	1.000	3	3
in the back	1.000	3	3
it should be	1.000	2	2
know if i	1.000	3	3
.action dances around	1.000	2	2
<post> oops sorry	1.000	2	2
las vegas </post>	1.000	2	2
bunch of people	1.000	3	3
the sake of	1.000	2	2
when you were	1.000	2	2
are you talking	1.000	2	2
ladies want to	1.000	3	3
<post> .action feels	1.000	6	6
<post> you just	1.000	4	4
just got a	1.000	2	2
you like the	1.000	2	2
lol me too	1.000	2	2
<post> a </post>	1.000	2	2
<post> u wish	1.000	2	2
have it. </post>	1.000	2	2
i just finished	1.000	2	2
i told my	1.000	2	2
<post> .action tosses	1.000	2	2
just like the	1.000	2	2
used to be	1.000	5	5
at you </post>	1.000	2	2
are going to	1.000	4	4
<post> so is	1.000	3	3
<post> well, i	1.000	3	3
i know you	1.000	5	5
to go and	1.000	2	2
<post> they had	1.000	2	2
to eat </post>	1.000	2	2
hey twisted </post>	1.000	2	2
he called me	1.000	2	2
<post> .action sneezes.	1.000	2	2

Teens Versus 30s			
Trigram	Entropy	30s	Teen
afk for a	1.000	2	2
have a good	1.000	12	12
<post> im an	1.000	2	2
<post> no idea	1.000	2	2
on fire </post>	1.000	2	2
<post> spin the	1.000	3	3
<post> will you	1.000	3	3
<post> any nice	1.000	2	2
<post> and your	1.000	3	3
the last thing	1.000	2	2
<post> how's that	1.000	2	2
ok lol </post>	1.000	2	2
<post> people are	1.000	2	2
here lol </post>	1.000	2	2
<post> but i'll	1.000	2	2
<post> so are	1.000	2	2
my friend </post>	1.000	2	2
<post> my profile	1.000	2	2
<post> lol is	1.000	3	3
<post> not </post>	1.000	3	3
trying to work	1.000	2	2
a slut </post>	1.000	2	2
would be </post>	1.000	2	2
you know how	1.000	3	3
<post> who </post>	1.000	3	3
i miss my	1.000	2	2
how to spell	1.000	3	3
<post> i also	1.000	2	2
<post> everyone is	1.000	2	2
seems to be	1.000	2	2
years ago </post>	1.000	3	3
<post> so do	1.000	4	4
ha ha ha	1.000	3	3
im out </post>	1.000	2	2
in it for	1.000	2	2
and that was	1.000	2	2
<post> but it's	1.000	4	4

Table 48. Entropy-Based Stop Trigrams for Teens vs. 30s Classification Task.

Teens Versus 40s			
Trigram	Entropy	40s	Teen
<post> you would	1.000	3	3
<post> wb gaston	1.000	2	2
<post> get the	1.000	2	2
it should be	1.000	2	2
<post> cause i	1.000	4	4
me up with	1.000	2	2
hey all </post>	1.000	8	8
<post> oops sorry	1.000	2	2
<post> u have	1.000	3	3
thought u were	1.000	2	2
to be with	1.000	2	2
he doesnt want	1.000	2	2
<post> sigh </post>	1.000	3	3
i just finished	1.000	2	2
<post> :-o </post>	1.000	7	7
just like the	1.000	2	2
for this room	1.000	2	2
<post> thats not	1.000	4	4
<post> is he	1.000	3	3
do you mean	1.000	2	2
<post> wb red	1.000	2	2
girl lol </post>	1.000	2	2
hey joe </post>	1.000	2	2
is back </post>	1.000	2	2
me ;-) </post>	1.000	2	2
i have another	1.000	2	2
<post> .action spanks	1.000	2	2
<post> it will	1.000	2	2
there was a	1.000	2	2
<post> talk to	1.000	6	6
a clue </post>	1.000	2	2
are you talkin	1.000	2	2
me from the	1.000	2	2
goes back to	1.000	2	2
i miss my	1.000	2	2
thought you were	1.000	6	6
don't know what	1.000	2	2
<post> i'm sorry	1.000	2	2

Teens Versus 40s			
Trigram	Entropy	40s	Teen
and i know	1.000	2	2
<post> tell me	1.000	4	4
trying to get	1.000	3	3
<post> .action cries.	1.000	3	3
that all the	1.000	2	2
what are u	1.000	3	3
<post> really </post>	1.000	2	2
it in </post>	1.000	2	2
seen it </post>	1.000	2	2
do you know	1.000	7	7
you in a	1.000	2	2
<post> ahhh </post>	1.000	2	2
prepare to be	1.000	2	2
<post> when you	1.000	2	2
<post> .action makes	1.000	3	3
is good </post>	1.000	7	7
<post> .action scratches	1.000	3	3
that was the	1.000	2	2
don't think i	1.000	3	3
hiya bob </post>	1.000	2	2
why do you	1.000	3	3
.action cries. </post>	1.000	3	3
<post> go for	1.000	2	2
:tongue: :tongue: </post>	1.000	3	3
thank you lol	1.000	2	2
u have to	1.000	4	4
yes they do	1.000	2	2
shakes her head	1.000	2	2
i was gonna	1.000	3	3
so i have	1.000	2	2
i get that	1.000	2	2
to take a	1.000	2	2
<post> oh that	1.000	2	2
if i had	1.000	5	5
<post> like a	1.000	3	3
<post> hey derby	1.000	3	3
give you a	1.000	2	2

Table 49. Entropy-Based Stop Trigrams for Teens vs. 40s Classification Task.

Teens Versus 50s			
Trigram	Entropy	50s	Teen
<post> what kind	1.000	3	3
wants me to	1.000	2	2
<post> did u	1.000	3	3
i have the	1.000	2	2
came back to	1.000	2	2
to get rid	1.000	2	2
i hope not	1.000	2	2
<post> why not?	1.000	2	2
.action giggles. </post>	1.000	2	2
if you have	1.000	2	2
you are going	1.000	2	2
<post> .action giggles.	1.000	2	2
it in the	1.000	2	2
<post> since when	1.000	3	3
i have an	1.000	2	2
<post> there is	1.000	3	3
have to get	1.000	2	2
a couple of	1.000	2	2
<post> same here	1.000	2	2
i'm back </post>	1.000	2	2
thats it </post>	1.000	2	2
back in a	1.000	3	3
you live in	1.000	2	2
i think that	1.000	2	2
you dont even	1.000	2	2
welcome back </post>	1.000	2	2
be back in	1.000	2	2
there is a	1.000	2	2
have a job	1.000	2	2
and i am	1.000	3	3
<post> what was	1.000	2	2
<post> lol ty	1.000	2	2
:-) :-) :-)	0.996	6	7
<post> hello room	0.991	4	5
so bored </post>	0.991	4	5
<post> hey there	0.991	4	5
i was a	0.991	4	5
:-) :-) </post>	0.985	4	3

Teens Versus 50s			
Trigram	Entropy	50s	Teen
<post> i always	0.985	3	4
in front of	0.985	4	3
pm me im	0.985	3	4
<post> ha ha	0.985	3	4
<post> i must	0.985	3	4
hello room </post>	0.971	4	6
<post> so is	0.971	2	3
hi everybody </post>	0.971	2	3
<post> have a	0.971	2	3
<post> am i	0.971	3	2
the world </post>	0.971	2	3
am back </post>	0.971	2	3
hiya dara </post>	0.971	3	2
<post> bye all	0.971	2	3
with you </post>	0.971	2	3
why do you	0.971	2	3
don't have to	0.971	2	3
i am in	0.971	3	2
<post> how long	0.971	2	3
you want to	0.971	2	3
i must be	0.971	2	3
<post> dont know	0.971	2	3
you are a	0.971	2	3
<post> hiya dara	0.971	3	2
thanks for the	0.971	2	3
<post> hi everybody	0.971	2	3
a long time	0.971	2	3
<post> must be	0.971	3	2
<post> wake up	0.971	2	3
do you think	0.971	2	3
to do with	0.971	2	3
<post> its ok	0.971	2	3
have you ever	0.971	3	2
hi joe </post>	0.971	2	3
dont even have	0.971	2	3
to hear that	0.971	2	3
bye all </post>	0.971	2	3

Table 50. Entropy-Based Stop Trigrams for Teens Versus 50s Classification Task.

Teens Versus Adults			
Trigram	Entropy	Adult	Teen
<post> get the	1.000	2	2
<post> .action cant	1.000	2	2
so bored </post>	1.000	5	5
trying to work	1.000	2	2
sup peeps </post>	1.000	2	2
<post> hey!! </post>	1.000	2	2
what you do	1.000	3	3
<post> need a	1.000	2	2
<post> i speak	1.000	2	2
i swear to	1.000	4	4
<post> whats that?	1.000	2	2
pretty cool </post>	1.000	2	2
you put in	1.000	2	2
hiya bob </post>	1.000	2	2
with the name	1.000	2	2
not for me	1.000	2	2
any girl from	1.000	2	2
he was in	1.000	2	2
<post> gross </post>	1.000	6	6
to what? </post>	1.000	2	2
<post> .action moans.	1.000	3	3
<post> hey brb	1.000	2	2
.action laughs and	1.000	2	2
<post> hey does	1.000	2	2
up in a	1.000	4	4
<post> ty. </post>	1.000	2	2
a night </post>	1.000	2	2
hell yea </post>	1.000	2	2
is everyone today?	1.000	2	2
in the mood	1.000	5	5
you calling me	1.000	2	2
wanna chat, pm	1.000	2	2
a go go	1.000	2	2
to stay </post>	1.000	2	2
sits back and	1.000	2	2
don't know who	1.000	3	3
my asl </post>	1.000	2	2
<post> because you're	1.000	2	2

Teens Versus Adults			
Trigram	Entropy	Adult	Teen
its not nice	1.000	2	2
<post> you'll get	1.000	2	2
goin on? </post>	1.000	2	2
ya i know	1.000	2	2
<post> you probably	1.000	2	2
<post> so yeah	1.000	3	3
a big ass	1.000	2	2
loves me </post>	1.000	5	5
<post> its really	1.000	2	2
<post> see u	1.000	2	2
never heard that	1.000	3	3
chat, pm me	1.000	2	2
that made me	1.000	2	2
bad lol </post>	1.000	2	2
stop talking about	1.000	3	3
i see. </post>	1.000	2	2
times a day	1.000	3	3
a lot to	1.000	2	2
know what? </post>	1.000	3	3
<post> i wear	1.000	4	4
no one want	1.000	2	2
to want to	1.000	2	2
hate u </post>	1.000	2	2
was it? </post>	1.000	2	2
havent talked to	1.000	2	2
do you? </post>	1.000	5	5
<post> .action pervs	1.000	2	2
will get you	1.000	2	2
not in this	1.000	2	2
big deal </post>	1.000	2	2
last thing i	1.000	2	2
.action takes out	1.000	2	2
<post> now you're	1.000	2	2
i seem to	1.000	3	3
hey holly </post>	1.000	6	6
i havent talked	1.000	2	2
<post> laffs </post>	1.000	2	2

Table 51. Entropy-Based Stop Trigrams for Teens Versus Adults Classification Task.

APPENDIX C: HIGH-FREQUENCY-BASED STOP WORD LISTS

This appendix contains the 75 n-grams in the high-frequency-based stop n-gram lists for the sixth random training set, which produced the best results for the teens versus adults classification task. This same set also performed well in the other classification tasks. Tables 52–66 contain the n-gram as well as its usage by the teens and the other age group in the classification task.

Teens Versus 20s		
Unigram	Teen	20s
.action	1511	2180
a	1828	4173
about	223	470
all	350	981
am	155	583
and	1184	2793
any	402	801
are	464	1256
at	255	610
be	398	925
but	443	866
can	243	638
chat	399	627
do	423	924
don't	187	469
dont	390	698
for	515	1237
from	219	580
get	324	753
go	267	577
good	233	805
got	210	446
have	529	1294
he	279	537
her	240	520
here	277	718
hey	740	1822
hi	360	1252
how	294	762
i	3459	7610
i'm	332	853
if	302	660
im	759	1202
in	855	2005
is	1029	2542
it	900	2041
its	298	601
just	398	929

Teens Versus 20s		
Unigram	Teen	20s
know	326	705
like	552	1159
lmao	242	681
lol	1987	5489
me	1206	2273
my	841	1910
no	485	1043
not	564	1288
of	573	1562
oh	236	472
ok	221	450
on	572	1286
one	261	608
or	236	517
out	227	553
pm	390	627
so	471	952
that	805	1921
the	1683	4162
they	218	469
think	177	447
this	246	515
to	1734	4129
too	230	524
u	630	1475
up	367	787
wanna	455	462
want	249	650
was	491	1014
we	151	443
well	209	475
what	388	985
with	482	1177
ya	143	442
yes	162	440
you	1594	3427
your	318	906

Table 52. Mutual High-Frequency Stop Unigrams for Teens Versus 20s Classification Task.

Teens Versus 30s		
Unigram	Teen	30s
.	240	304
.action	1511	1307
a	1828	2377
about	223	286
all	350	583
an	146	259
and	1184	1743
are	464	567
at	255	375
back	202	264
be	398	472
but	443	494
can	243	317
did	152	254
do	423	481
don't	187	395
for	515	810
from	219	304
get	324	440
go	267	313
good	233	416
got	210	268
haha	129	379
have	529	778
he	279	382
hello	107	356
her	240	316
here	277	338
hey	740	1015
hi	360	757
how	294	366
i	3459	3816
i'm	332	459
if	302	385
im	759	391
in	855	1353
is	1029	1440
it	900	1063

Teens Versus 30s		
Unigram	Teen	30s
it's	127	259
just	398	449
know	326	354
like	552	625
lmao	242	368
lol	1987	2304
me	1206	896
my	841	1091
no	485	499
not	564	686
of	573	993
oh	236	400
ok	221	344
on	572	749
one	261	352
or	236	309
out	227	336
so	471	487
some	183	264
that	805	1289
the	1683	2683
they	218	398
think	177	281
this	246	301
to	1734	2258
too	230	388
u	630	463
up	367	384
was	491	654
wb	144	341
we	151	270
well	209	323
what	388	504
with	482	533
ya	143	315
you	1594	1816
your	318	400

Table 53. Mutual High-Frequency Stop Unigrams for Teens Versus 30s Classification Task.

Teens Versus 40s		
Unigram	Teen	40s
.action	1511	861
:)	158	597
a	1828	1712
about	223	180
all	350	487
am	155	247
and	1184	1208
are	464	527
at	255	310
back	202	232
be	398	411
but	443	280
can	243	282
did	152	196
do	423	350
don't	187	259
for	515	594
from	219	175
get	324	253
go	267	194
good	233	384
have	529	494
he	279	295
hello	107	267
her	240	267
here	277	405
hey	740	1422
hi	360	2576
hiya	269	292
how	294	275
i	3459	2665
i'm	332	294
if	302	223
in	855	949
is	1029	1227
it	900	807
just	398	436
know	326	314

Teens Versus 40s		
Unigram	Teen	40s
like	552	409
lol	1987	4479
love	217	179
me	1206	737
my	841	671
no	485	404
not	564	501
now	182	178
of	573	630
oh	236	233
ok	221	284
on	572	490
one	261	214
or	236	182
out	227	226
see	151	227
she	192	209
so	471	290
that	805	916
the	1683	1851
there	216	284
they	218	222
this	246	193
to	1734	1548
too	230	359
u	630	259
up	367	257
was	491	491
wb	144	493
we	151	224
well	209	201
what	388	427
with	482	338
ya	143	281
yes	162	183
you	1594	1587
your	318	337

Table 54. Mutual High-Frequency Stop Unigrams for Teens Versus 40s Classification Task.

Teens Versus 50s		
Unigram	Teen	50s
a	1828	309
all	350	75
am	155	35
an	146	34
and	1184	201
any	402	55
are	464	66
as	131	30
at	255	34
back	202	50
be	398	67
but	443	41
can	243	79
chat	399	62
did	152	34
do	423	41
don't	187	32
for	515	105
from	219	33
get	324	39
go	267	37
good	233	40
got	210	34
has	115	29
have	529	94
he	279	34
hello	107	59
here	277	55
hey	740	101
hi	360	309
his	110	29
hiya	269	87
how	294	37
i	3459	394
i'm	332	40
if	302	37
in	855	191
is	1029	158

Teens Versus 50s		
Unigram	Teen	50s
it	900	108
its	298	30
just	398	52
like	552	53
lmao	242	38
lol	1987	331
love	217	30
me	1206	135
my	841	111
no	485	51
not	564	76
now	182	29
of	573	122
on	572	75
or	236	30
out	227	47
pm	390	62
room	116	29
so	471	44
that	805	117
the	1683	317
there	216	30
they	218	42
think	177	35
to	1734	265
too	230	37
u	630	29
up	367	35
want	249	52
was	491	70
we	151	30
what	388	42
when	163	36
where	107	30
with	482	66
you	1594	219
your	318	45

Table 55. Mutual High-Frequency Stop Unigrams for Teens Versus 50s Classification Task.

Teens Versus Adults		
Unigram	Teen	Adult
.action	1511	5038
a	1828	8139
about	223	997
all	350	1980
am	155	951
and	1184	5770
any	402	1090
are	464	2381
at	255	1275
be	398	1879
but	443	1599
can	243	1311
do	423	1673
don't	187	988
dont	390	1107
for	515	2615
from	219	1050
get	324	1480
go	267	1119
good	233	1517
got	210	872
have	529	2472
he	279	1209
hello	107	1070
her	240	1064
here	277	1376
hey	740	4147
hi	360	4770
how	294	1518
i	3459	13921
i'm	332	1540
if	302	1374
im	759	1818
in	855	4339
is	1029	5692
it	900	4014
its	298	1022
just	398	1817

Teens Versus Adults		
Unigram	Teen	Adult
know	326	1348
like	552	2148
lmao	242	871
lol	1987	10532
me	1206	3934
my	841	3725
no	485	1976
not	564	2482
of	573	3165
oh	236	1169
ok	221	1099
on	572	2618
one	261	1162
or	236	1006
out	227	1149
pm	390	936
so	471	1719
that	805	3919
the	1683	8788
there	216	925
they	218	1047
think	177	881
this	246	1063
to	1734	8531
too	230	1069
u	630	2057
up	367	1430
want	249	995
was	491	2080
wb	144	1051
we	151	927
well	209	1033
what	388	1918
with	482	2127
ya	143	1048
you	1594	7002
your	318	1708

Table 56. Mutual High-Frequency Stop Unigrams for Teens Versus Adults Classification Task.

Teens Versus 20s		
Bigram	Teen	20s
.action is	63	196
:) </post>	150	320
<post> !seen	97	189
<post> .action	1511	2180
<post> and	262	509
<post> any	248	534
<post> but	189	311
<post> bye	81	235
<post> good	63	234
<post> haha	119	209
<post> hello	90	365
<post> hey	692	1760
<post> hi	323	1189
<post> how	146	381
<post> i	1651	3468
<post> i'm	152	429
<post> im	381	640
<post> it	132	233
<post> its	148	290
<post> lmao	209	653
<post> lol	1546	4847
<post> my	150	258
<post> no	269	590
<post> not	116	248
<post> oh	215	414
<post> ok	161	268
<post> so	149	320
<post> that	104	258
<post> thats	132	202
<post> the	87	238
<post> u	102	262
<post> wb	137	265
<post> well	165	332
<post> what	171	407
<post> whats	82	183
<post> why	72	184
<post> yeah	164	279
<post> yes	128	361

Teens Versus 20s		
Bigram	Teen	20s
<post> you	271	527
? </post>	52	239
all </post>	74	228
and i	93	194
are you	108	244
chat with	69	234
have a	105	273
have to	87	198
here </post>	98	326
hi </post>	77	204
i am	96	447
i can	69	183
i don't	81	227
i dont	184	290
i have	152	401
i know	111	249
i like	95	186
i love	123	200
i think	85	222
i was	171	328
in the	173	394
is a	71	208
it </post>	152	360
lmao </post>	133	381
lol </post>	1424	3827
me </post>	452	833
of the	60	198
on the	82	221
pm me	320	457
that </post>	75	190
to be	100	221
to chat	96	274
to the	55	222
wanna chat	226	208
want to	145	398
with a	79	268
you </post>	130	204

Table 57. Mutual High-Frequency Stop Bigrams for Teens Versus 20s Classification Task.

Teens Versus 30s		
Bigram	Teen	30s
.action is	63	179
:) </post>	150	131
<post> .action	1511	1307
<post> and	262	271
<post> any	248	84
<post> but	189	127
<post> bye	81	85
<post> good	63	118
<post> haha	119	164
<post> he	89	102
<post> hello	90	346
<post> hey	692	973
<post> hi	323	728
<post> how	146	192
<post> i	1651	1519
<post> i'm	152	164
<post> im	381	205
<post> is	95	89
<post> it	132	107
<post> lmao	209	351
<post> lol	1546	1751
<post> my	150	158
<post> no	269	240
<post> not	116	150
<post> oh	215	377
<post> ok	161	231
<post> so	149	168
<post> that	104	106
<post> the	87	107
<post> ty	62	125
<post> wb	137	337
<post> well	165	178
<post> what	171	184
<post> yeah	164	191
<post> yes	128	124
<post> you	271	280
all </post>	74	145
and i	93	101

Teens Versus 30s		
Bigram	Teen	30s
are you	108	99
but i	81	91
for a	59	81
haha </post>	89	369
have a	105	146
have to	87	138
here </post>	98	120
i am	96	151
i can	69	85
i don't	81	188
i dont	184	101
i have	152	212
i just	77	83
i know	111	111
i like	95	89
i think	85	165
i was	171	173
if you	69	89
in a	70	96
in the	173	271
is a	71	106
it </post>	152	156
it was	54	93
lmao </post>	133	201
lol </post>	1424	1423
me </post>	452	178
now </post>	58	86
of the	60	150
ok </post>	124	94
on the	82	147
that </post>	75	97
to be	100	91
to the	55	139
too </post>	74	121
want to	145	85
with a	79	85
you </post>	130	106

Table 58. Mutual High-Frequency Stop Bigrams for Teens Versus 30s Classification Task.

Teens Versus 40s		
Bigram	Teen	40s
. </post>	186	112
.action is	63	97
:) </post>	150	504
<post> .action	1511	861
<post> and	262	154
<post> but	189	74
<post> bye	81	104
<post> gm	56	226
<post> good	63	133
<post> he	89	74
<post> hello	90	212
<post> hey	692	1379
<post> hi	323	2519
<post> hiya	263	279
<post> how	146	129
<post> i	1651	1021
<post> i'm	152	152
<post> is	95	135
<post> it	132	90
<post> it's	68	65
<post> lmao	209	97
<post> lol	1546	2463
<post> me	65	85
<post> my	150	77
<post> no	269	217
<post> not	116	97
<post> oh	215	212
<post> ok	161	181
<post> omg	83	106
<post> so	149	77
<post> that	104	96
<post> thats	132	77
<post> the	87	109
<post> ty	62	187
<post> wb	137	430
<post> well	165	134
<post> what	171	172
<post> yeah	164	104

Teens Versus 40s		
Bigram	Teen	40s
<post> yes	128	143
<post> you	271	259
all </post>	74	124
and i	93	66
are you	108	107
back </post>	58	76
do you	74	74
have a	105	118
have to	87	67
here </post>	98	114
i am	96	190
i can	69	62
i don't	81	104
i have	152	131
i know	111	84
i like	95	69
i think	85	82
i was	171	135
in a	70	75
in the	173	212
is a	71	87
it </post>	152	107
it is	53	73
it was	54	78
lmao </post>	133	72
lol </post>	1424	2834
lol. </post>	65	95
me </post>	452	106
of the	60	85
on the	82	84
that </post>	75	75
to be	100	93
too </post>	74	101
want to	145	70
you </post>	130	77
you are	58	76
you? </post>	60	67

Table 59. Mutual High-Frequency Stop Bigrams for Teens Versus 40s Classification Task.

Teens Versus 50s		
Bigram	Teen	50s
:) </post>	150	15
<post> .action	1511	23
<post> and	262	27
<post> any	248	42
<post> but	189	8
<post> gm	56	31
<post> good	63	13
<post> hello	90	53
<post> hey	692	97
<post> hi	323	307
<post> hiya	263	79
<post> how	146	17
<post> i	1651	127
<post> i'm	152	13
<post> im	381	8
<post> is	95	26
<post> it	132	17
<post> lmao	209	37
<post> lol	1546	258
<post> my	150	12
<post> no	269	20
<post> not	116	17
<post> oh	215	23
<post> ok	161	11
<post> omg	83	14
<post> so	149	9
<post> the	87	15
<post> they	58	8
<post> ty	62	28
<post> u	102	10
<post> wb	137	27
<post> well	165	20
<post> what	171	15
<post> who	91	11
<post> yes	128	19
<post> you	271	20
? </post>	52	21
all </post>	74	28

Teens Versus 50s		
Bigram	Teen	50s
and i	93	10
are you	108	8
back </post>	58	19
chat pm	79	9
for a	59	18
going to	57	8
have a	105	26
have to	87	14
here </post>	98	12
i am	96	29
i don't	81	11
i have	152	23
i love	123	8
i think	85	19
i was	171	19
in a	70	18
in the	173	30
is a	71	13
it </post>	152	10
it is	53	10
lol </post>	1424	150
me </post>	452	63
now </post>	58	12
of the	60	16
on the	82	14
pm me	320	50
that </post>	75	13
to be	100	17
to chat	96	11
to the	55	11
too </post>	74	11
want to	145	16
with a	79	9
you </post>	130	27
you are	58	13
you have	54	8
you? </post>	60	8

Table 60. Mutual High-Frequency Stop Bigrams for Teens Versus 50s Classification Task.

Teens Versus Adults		
Bigram	Teen	Adult
.action is	63	995
:) </post>	150	758
<post> .action	1511	5038
<post> and	262	893
<post> any	248	639
<post> but	189	551
<post> bye	81	384
<post> good	63	416
<post> he	89	335
<post> hello	90	997
<post> hey	692	4006
<post> hi	323	4641
<post> hiya	263	462
<post> how	146	742
<post> i	1651	5985
<post> i'm	152	739
<post> if	91	320
<post> im	381	886
<post> is	95	345
<post> it	132	471
<post> its	148	427
<post> lmao	209	802
<post> lol	1546	8219
<post> my	150	485
<post> no	269	1080
<post> not	116	500
<post> oh	215	1059
<post> ok	161	674
<post> so	149	525
<post> that	104	454
<post> thats	132	367
<post> the	87	468
<post> ty	62	478
<post> u	102	322
<post> wb	137	988
<post> well	165	685
<post> what	171	770
<post> yeah	164	591

Teens Versus Adults		
Bigram	Teen	Adult
<post> yes	128	559
<post> you	271	1010
? </post>	52	315
all </post>	74	488
and i	93	354
are you	108	499
have a	105	517
have to	87	389
here </post>	98	479
i am	96	728
i can	69	350
i don't	81	404
i dont	184	473
i have	152	681
i know	111	435
i like	95	334
i love	123	342
i think	85	460
i was	171	647
in the	173	864
is a	71	404
it </post>	152	664
lmao </post>	133	540
lol </post>	1424	7035
me </post>	452	1158
of the	60	405
ok </post>	124	334
on the	82	438
pm me	320	580
that </post>	75	364
to be	100	418
to chat	96	364
to the	55	431
too </post>	74	363
want to	145	538
with a	79	417
you </post>	130	436

Table 61. Mutual High-Frequency Stop Bigrams for Teens Versus Adults Classification Task.

Teens Versus 20s		
Trigram	Teen	Adult
<post> !scramble </post>	6	87
<post> .action is	63	196
<post> .action looks	15	66
<post> :) </post>	57	84
<post> :-o </post>	8	75
<post> :o </post>	28	60
<post> ;) </post>	8	66
<post> ? </post>	34	55
<post> and i	35	61
<post> any ladies	17	273
<post> brb </post>	46	100
<post> but i	34	59
<post> do you	19	48
<post> haha </post>	81	140
<post> hahaha </post>	23	72
<post> hehe </post>	19	54
<post> hello </post>	26	105
<post> hey </post>	62	92
<post> hey all	17	46
<post> hey everyone	21	52
<post> hey room	22	47
<post> hi </post>	69	181
<post> how are	27	111
<post> i am	48	206
<post> i can	23	65
<post> i do	26	55
<post> i don't	42	116
<post> i dont	92	145
<post> i got	32	68
<post> i hate	41	62
<post> i have	80	214
<post> i just	39	83
<post> i know	60	124
<post> i like	67	116
<post> i love	86	121
<post> i need	26	78
<post> i think	49	119
<post> i want	32	80

Teens Versus 20s		
Trigram	Teen	Adult
<post> i was	67	131
<post> i'm not	15	55
<post> im not	39	64
<post> it was	20	53
<post> lmao </post>	105	360
<post> lol </post>	1042	3236
<post> lol i	35	83
<post> lol! </post>	12	67
<post> lol. </post>	35	58
<post> me too	18	48
<post> no </post>	51	107
<post> oh </post>	30	47
<post> ok </post>	90	82
<post> that is	6	52
<post> well i	17	47
<post> whats up	29	66
<post> wow </post>	33	46
<post> yeah </post>	53	71
<post> yeah i	13	47
<post> yes </post>	42	105
<post> yw </post>	35	51
any ladies wanna	9	94
chat pm me	76	107
chat with a	33	181
hey everyone </post>	20	46
how are you	15	73
i have a	43	91
i have to	27	61
i know </post>	28	47
i want to	18	59
in here </post>	24	48
ladies wanna chat	11	76
pm me </post>	225	348
to chat with	28	170
to me </post>	25	60
wanna chat pm	49	89
want to chat	62	196

Table 62. Mutual High-Frequency Stop Trigrams for Teens Versus 20s Classification Task.

Teens Versus 30s		
Trigram	Teen	Adult
<post> .action gets	16	24
<post> .action gives	25	22
<post> .action has	19	26
<post> .action is	63	179
<post> .action looks	15	33
<post> .action sits	28	48
<post> :) </post>	57	28
<post> and i	35	31
<post> any ladies	17	28
<post> brb </post>	46	41
<post> but i	34	25
<post> cool </post>	14	29
<post> did you	12	23
<post> haha </post>	81	155
<post> hahaha </post>	23	54
<post> hello </post>	26	30
<post> hey </post>	62	33
<post> hi </post>	69	39
<post> hi all	13	50
<post> how are	27	36
<post> i am	48	72
<post> i can	23	26
<post> i do	26	26
<post> i don't	42	63
<post> i dont	92	50
<post> i dunno	17	33
<post> i got	32	22
<post> i had	15	30
<post> i hate	41	26
<post> i have	80	74
<post> i just	39	31
<post> i know	60	43
<post> i like	67	41
<post> i love	86	40
<post> i need	26	21
<post> i think	49	84
<post> i thought	32	20
<post> i want	32	25

Teens Versus 30s		
Trigram	Teen	Adult
<post> i was	67	58
<post> i'm not	15	21
<post> if i	18	20
<post> it is	12	22
<post> it was	20	20
<post> lmao </post>	105	186
<post> lmfao </post>	30	114
<post> lol </post>	1042	909
<post> lol i	35	26
<post> me too	18	27
<post> no i	20	23
<post> oh </post>	30	28
<post> ok </post>	90	38
<post> omg </post>	41	21
<post> that was	22	25
<post> ty </post>	19	24
<post> well i	17	29
<post> what is	16	23
<post> yea </post>	44	21
<post> yeah i	13	20
<post> yep </post>	19	27
<post> yes </post>	42	37
<post> you are	12	20
<post> you know	15	22
<post> yw </post>	35	68
a lot of	9	21
hi all </post>	12	38
i don't know	21	32
i had a	10	20
i have a	43	55
i have to	27	41
i know </post>	28	20
i need to	21	22
i think i	19	34
i used to	10	20
pm me </post>	225	27
you have to	15	24

Table 63. Mutual High-Frequency Stop Trigrams for Teens Versus 30s Classification Task.

Teens Versus 40s		
Trigram	Teen	Adult
<post> .action has	19	43
<post> .action is	63	97
<post> .action looks	15	52
<post> .action sits	28	19
<post> :) </post>	57	38
<post> :-) </post>	29	43
<post> ;-) </post>	23	14
<post> ? </post>	34	15
<post> and i	35	17
<post> are you	31	18
<post> back </post>	10	19
<post> brb </post>	46	45
<post> can i	9	13
<post> did you	12	14
<post> do you	19	18
<post> hello </post>	26	21
<post> hi </post>	69	30
<post> hi all	13	44
<post> how are	27	27
<post> i am	48	82
<post> i can	23	18
<post> i do	26	15
<post> i don't	42	44
<post> i dont	92	17
<post> i had	15	15
<post> i have	80	51
<post> i just	39	28
<post> i know	60	39
<post> i like	67	38
<post> i love	86	38
<post> i need	26	18
<post> i see	19	16
<post> i think	49	28
<post> i wanna	26	15
<post> i want	32	14
<post> i was	67	54
<post> i would	17	25
<post> i'm not	15	17

Teens Versus 40s		
Trigram	Teen	Adult
<post> is that	18	21
<post> it is	12	20
<post> it was	20	25
<post> lmao </post>	105	44
<post> lol </post>	1042	933
<post> lol @	32	52
<post> me too	18	50
<post> no </post>	51	16
<post> ok </post>	90	30
<post> omg </post>	41	15
<post> oops </post>	11	16
<post> thank you	13	19
<post> this is	15	15
<post> ty </post>	19	18
<post> well i	17	18
<post> what is	16	20
<post> who is	11	14
<post> yep </post>	19	24
<post> yes </post>	42	18
<post> you are	12	14
<post> you can	8	15
<post> you have	12	15
<post> you know	15	23
are you? </post>	21	17
for me </post>	9	13
hi all </post>	12	44
how are you	15	21
i am not	13	20
i don't know	21	15
i have a	43	25
i have to	27	18
i know </post>	28	14
i need to	21	13
i want to	18	15
me too </post>	15	14
what do you	9	18
you have a	9	14

Table 64. Mutual High-Frequency Stop Trigrams for Teens Versus 40s Classification Task.

Teens Versus 50s		
Trigram	Teen	Adult
:love: :love: </post>	12	5
<post> .action is	63	4
<post> :-) </post>	29	5
<post> ;-) </post>	23	4
<post> ? </post>	34	3
<post> and i	35	3
<post> any ladies	17	34
<post> back </post>	10	5
<post> brb </post>	46	4
<post> hello </post>	26	3
<post> hey </post>	62	3
<post> hey room	22	3
<post> hi </post>	69	4
<post> hi all	13	4
<post> how are	27	4
<post> i am	48	4
<post> i did	10	2
<post> i do	26	3
<post> i dunno	17	2
<post> i got	32	2
<post> i have	80	8
<post> i just	39	4
<post> i know	60	2
<post> i like	67	4
<post> i love	86	4
<post> i saw	8	2
<post> i see	19	2
<post> i think	49	11
<post> i was	67	6
<post> i would	17	2
<post> i'm so	8	2
<post> im a	19	2
<post> is that	18	3
<post> it was	20	4
<post> lmao </post>	105	4
<post> lol </post>	1042	79
<post> lol @	32	5
<post> lol i	35	3

Teens Versus 50s		
Trigram	Teen	Adult
<post> me too	18	2
<post> of course	9	2
<post> oh ok	10	2
<post> ok </post>	90	2
<post> omg </post>	41	2
<post> sorry </post>	12	2
<post> that was	22	3
<post> ty </post>	19	2
<post> well i	17	6
<post> wow </post>	33	4
<post> yeah i	13	2
<post> you are	12	3
<post> you know	15	4
<post> yw </post>	35	4
a lot of	9	4
chat pm me	76	9
for me to	9	2
have a good	9	2
hello everyone </post>	8	2
hey room </post>	16	3
hi all </post>	12	3
how are you	15	4
i don't know	21	3
i got a	11	2
i have a	43	6
i have to	27	2
i like the	10	2
i love the	11	2
i think i	19	3
i thought you	10	2
i used to	10	3
i want to	18	2
in here </post>	24	2
know how to	16	2
pm me </post>	225	45
to me </post>	25	3
want to chat	62	5

Table 65. Mutual High-Frequency Stop Trigrams for Teens Versus 50s Classification Task.

Teens Versus Adults		
Trigram	Teen	Adult
.action looks at	6	96
<post> !scramble </post>	6	354
<post> .action is	63	995
<post> .action looks	15	169
<post> .action sits	28	125
<post> :) </post>	57	142
<post> :-) </post>	29	84
<post> and i	35	97
<post> any ladies	17	280
<post> are you	31	96
<post> brb </post>	46	180
<post> but i	34	99
<post> haha </post>	81	193
<post> hahaha </post>	23	137
<post> hello </post>	26	156
<post> hello all	7	102
<post> hey </post>	62	148
<post> hi </post>	69	208
<post> hi all	13	117
<post> how are	27	204
<post> how is	8	81
<post> i am	48	331
<post> i can	23	112
<post> i do	26	78
<post> i don't	42	183
<post> i dont	92	232
<post> i got	32	102
<post> i had	15	86
<post> i hate	41	113
<post> i have	80	315
<post> i just	39	148
<post> i know	60	214
<post> i like	67	197
<post> i love	86	199
<post> i need	26	114
<post> i think	49	242
<post> i want	32	102
<post> i was	67	241

Teens Versus Adults		
Trigram	Teen	Adult
<post> i'm not	15	90
<post> im not	39	97
<post> it was	20	97
<post> lmao </post>	105	477
<post> lmfao </post>	30	155
<post> lol </post>	1042	4866
<post> lol @	32	128
<post> lol i	35	110
<post> lol. </post>	35	98
<post> me too	18	90
<post> no </post>	51	134
<post> ok </post>	90	140
<post> that is	6	81
<post> that was	22	86
<post> ty </post>	19	103
<post> well i	17	98
<post> what is	16	84
<post> whats up	29	88
<post> yeah </post>	53	94
<post> yeah i	13	85
<post> yep </post>	19	95
<post> yes </post>	42	143
<post> you know	15	83
<post> yw </post>	35	146
are you? </post>	21	79
chat pm me	76	80
chat with a	33	195
hi all </post>	12	101
how are you	15	134
i have a	43	151
i have to	27	118
i know </post>	28	80
i think i	19	77
pm me </post>	225	403
to chat with	28	188
to me </post>	25	78
want to chat	62	214

Table 66. Mutual High-Frequency Stop Trigrams for Teens Versus Adults Classification Task.

APPENDIX D: NAÏVE BAYES CLASSIFIER RESULTS

This appendix contains the Naïve Bayes Classifier results. The results for each classification task are ranked by average F-score. In order to exclude outliers, the average F-score was calculated without the highest and lowest F-score measure from the 10 random test sets. The lowest and highest F-score results, however, are included in the tables as separate columns. The average number of true positives, false positives, true negatives, false negatives, precision and recall do not omit the highest/lowest respective value from the 10 random test set results. Tables 67–70 display the Naïve Bayes Classifier results for each classification task.

Classification Task	Feature	Teen Training Files	Adult Training Files	True Positives	False Positives	False Negatives	True Negatives	Precision	Recall	Low F-Score	High F-Score	Average F-Score
Teens vs. 20s	Trigram	465	689	104.2	77.9	11.8	94.1	0.567	0.898	0.438	0.766	0.717
	3 Character Gram	465	689	57.6	56.4	58.4	115.6	0.439	0.497	0.054	0.818	0.466
	Unigram	465	689	53.5	45.7	62.5	126.3	0.482	0.461	0.105	0.827	0.462
	4 Character Gram	465	689	47.7	36.7	68.3	135.3	0.488	0.411	0.026	0.866	0.433
	5 Character Gram	465	689	27.6	32.3	88.4	139.7	0.390	0.238	0.000	0.629	0.285
	Bigram	465	689	20.2	45.3	95.8	126.7	0.278	0.174	0.038	0.419	0.207
Teens vs. 30s	3 Character Gram	465	259	115.3	28.1	0.7	36.9	0.804	0.994	0.879	0.899	0.889
	Bigram	465	259	109.9	28.6	6.1	36.4	0.790	0.947	0.627	0.903	0.884
	4 Character Gram	465	259	108.1	22.8	7.9	42.2	0.825	0.932	0.774	0.924	0.880
	5 Character Gram	465	259	102.9	20.2	13.1	44.8	0.830	0.887	0.561	0.947	0.879
	Unigram	465	259	110.6	30.4	5.4	34.6	0.783	0.953	0.707	0.896	0.873
	Trigram	465	259	114.7	37.1	1.3	27.9	0.756	0.989	0.844	0.869	0.857
Teens vs. 40s	4 Character Gram	465	235	114.4	9.9	1.6	49.1	0.921	0.986	0.921	0.970	0.954
	5 Character Gram	465	235	114.3	10.5	1.7	48.5	0.916	0.985	0.904	0.967	0.953
	3 Character Gram	465	235	114.7	12.6	1.3	46.4	0.901	0.989	0.915	0.963	0.944
	Unigram	465	235	114.2	12.3	1.8	46.7	0.903	0.984	0.890	0.979	0.944
	Bigram	465	235	115.2	15.2	0.8	43.8	0.884	0.993	0.905	0.954	0.937
	Trigram	465	235	115	16.2	1	42.8	0.877	0.991	0.900	0.963	0.930
Teens vs. 50s	Trigram	465	80	114.6	4.7	1.4	15.3	0.961	0.988	0.957	0.987	0.975
	3 Character Gram	465	80	116	11.4	0	8.6	0.911	1.000	0.939	0.967	0.953
	Unigram	465	80	115.9	12.3	0.1	7.7	0.904	0.999	0.939	0.955	0.950
	Bigram	465	80	115.9	13.1	0.1	6.9	0.899	0.999	0.935	0.963	0.945
	4 Character Gram	465	80	116	13.6	0	6.4	0.895	1.000	0.932	0.959	0.945
	5 Character Gram	465	80	115.9	15.7	0.1	4.3	0.881	0.999	0.924	0.947	0.936
Teens vs. Adults	Trigram	465	1263	79.3	110.2	36.7	205.8	0.406	0.684	0.245	0.699	0.516
	3 Character Gram	465	1263	44.2	60.8	71.8	255.2	0.364	0.381	0.022	0.730	0.363
	Unigram	465	1263	35.3	50.7	80.7	265.3	0.365	0.304	0.000	0.814	0.305
	4 Character Gram	465	1263	22	41.5	94	274.5	0.269	0.190	0.000	0.600	0.199
	Bigram	465	1263	15.5	53.2	100.5	262.8	0.205	0.134	0.000	0.347	0.157
	5 Character Gram	465	1263	7.7	31.6	108.3	284.4	0.170	0.066	0.000	0.275	0.083

Table 67. Naïve Bayes Classifier Results Ranked by Average F-score for Each Classification Task (Whitten-Bell Smoothing with Punctuation).

Classification Task	Feature	Teen Training Files	Adult Training Files	True Positives	False Positives	False Negatives	True Negatives	Precision	Recall	Low F-Score	High F-Score	Average F-Score
Teens vs. 20s	Trigram	465	689	104.7	77.1	11.3	94.9	0.559	0.903	0.194	0.764	0.741
	3 Character Gram	465	689	56	56.2	60	115.8	0.424	0.483	0.011	0.824	0.452
	4 Character Gram	465	689	44.2	35.7	71.8	136.3	0.462	0.381	0.025	0.871	0.394
	5 Character Gram	465	689	31.2	30.3	84.8	141.7	0.398	0.269	0.000	0.752	0.295
	Unigram	465	689	27.4	45.8	88.6	126.2	0.335	0.236	0.035	0.611	0.261
	Bigram	465	689	19.9	45.6	96.1	126.4	0.283	0.172	0.061	0.383	0.210
Teens vs. 30s	5 Character Gram	465	259	104.5	20	11.5	45	0.834	0.901	0.564	0.939	0.889
	4 Character Gram	465	259	109.2	22.6	6.8	42.4	0.829	0.941	0.771	0.921	0.889
	3 Character Gram	465	259	113.3	27.7	2.7	37.3	0.804	0.977	0.850	0.903	0.883
	Bigram	465	259	109	32.4	7	32.6	0.765	0.940	0.531	0.885	0.873
	Unigram	465	259	108.1	29.8	7.9	35.2	0.780	0.932	0.587	0.899	0.872
	Trigram	465	259	114.9	38.5	1.1	26.5	0.749	0.991	0.843	0.867	0.853
Teens vs. 40s	5 Character Gram	465	235	114.3	9.7	1.7	49.3	0.922	0.985	0.904	0.971	0.956
	4 Character Gram	465	235	114.3	11.1	1.7	47.9	0.912	0.985	0.907	0.967	0.949
	Unigram	465	235	114.1	10.7	1.9	48.3	0.915	0.984	0.908	0.983	0.948
	3 Character Gram	465	235	114.1	15	1.9	44	0.884	0.984	0.874	0.947	0.936
	Bigram	465	235	115.4	17.3	0.6	41.7	0.870	0.995	0.899	0.950	0.929
	Trigram	465	235	115.2	22.1	0.8	36.9	0.840	0.993	0.887	0.939	0.909
Teens vs. 50s	Trigram	465	80	114.6	5.6	1.4	14.4	0.953	0.988	0.952	0.979	0.971
	Unigram	465	80	115.9	12.7	0.1	7.3	0.901	0.999	0.939	0.955	0.948
	3 Character Gram	465	80	116	13.2	0	6.8	0.898	1.000	0.935	0.955	0.946
	Bigram	465	80	115.9	13.8	0.1	6.2	0.894	0.999	0.935	0.955	0.943
	4 Character Gram	465	80	116	14.1	0	5.9	0.892	1.000	0.932	0.955	0.943
	5 Character Gram	465	80	116	14.6	0	5.4	0.888	1.000	0.928	0.951	0.941
Teens vs. Adults	Trigram	465	1263	98.6	113.1	17.4	202.9	0.452	0.850	0.141	0.712	0.630
	3 Character Gram	465	1263	46.2	58.5	69.8	257.5	0.371	0.398	0.000	0.765	0.376
	4 Character Gram	465	1263	30	37.6	86	278.4	0.316	0.259	0.000	0.744	0.251
	Bigram	465	1263	18.4	53.7	97.6	262.3	0.238	0.159	0.021	0.366	0.187
	5 Character Gram	465	1263	15.7	30.9	100.3	285.1	0.277	0.135	0.000	0.610	0.145
	Unigram	465	1263	14.4	46.6	101.6	269.4	0.201	0.124	0.000	0.376	0.143

Table 68. Naïve Bayes Classifier Results Ranked by Average F-score for Each Classification Task (Whitten-Bell Smoothing without Punctuation).

Classification Task	Feature	Teen Training Files	Adult Training Files	True Positives	False Positives	False Negatives	True Negatives	Precision	Recall	Low F-Score	High F-Score	Average F-Score
Teens vs. 20s	Trigram	465	689	2.6	12.4	113.4	159.6	0.114	0.022	0.000	0.240	0.015
	Bigram	465	689	0.3	3.5	115.7	168.5	0.073	0.003	0.000	0.017	0.004
	3 Character Gram	465	689	0	0	116	172	0.000	0.000	0.000	0.000	0.000
	4 Character Gram	465	689	0	0.2	116	171.8	0.000	0.000	0.000	0.000	0.000
	5 Character Gram	465	689	0	0.9	116	171.1	0.000	0.000	0.000	0.000	0.000
	Unigram	465	689	0	0.1	116	171.9	0.000	0.000	0.000	0.000	0.000
Teens vs. 30s	Trigram	465	259	105.8	28.1	10.2	36.9	0.784	0.912	0.520	0.910	0.869
	Bigram	465	259	77.1	25.2	38.9	39.8	0.699	0.665	0.144	0.889	0.691
	5 Character Gram	465	259	19	7.1	97	57.9	0.513	0.164	0.000	0.779	0.183
	Unigram	465	259	19.9	10.7	96.1	54.3	0.406	0.172	0.000	0.839	0.163
	4 Character Gram	465	259	6.4	4.7	109.6	60.3	0.369	0.055	0.000	0.325	0.073
	3 Character Gram	465	259	2.1	1.5	113.9	63.5	0.185	0.018	0.000	0.167	0.020
Teens vs. 40s	Unigram	465	235	107.6	11.1	8.4	47.9	0.906	0.928	0.548	0.967	0.943
	5 Character Gram	465	235	104.8	8.7	11.2	50.3	0.920	0.903	0.475	0.971	0.941
	Bigram	465	235	115.6	20.5	0.4	38.5	0.850	0.997	0.906	0.935	0.916
	Trigram	465	235	115.5	24.3	0.5	34.7	0.827	0.996	0.885	0.921	0.903
	4 Character Gram	465	235	92.4	7.3	23.6	51.7	0.852	0.797	0.033	0.975	0.855
	3 Character Gram	465	235	86.6	7.9	29.4	51.1	0.864	0.747	0.017	0.979	0.796
Teens vs. 50s	Trigram	465	80	115.9	19.3	0.1	0.7	0.857	0.999	0.916	0.928	0.923
	5 Character Gram	465	80	116	19.3	0	0.7	0.857	1.000	0.921	0.928	0.923
	Bigram	465	80	116	19.7	0	0.3	0.855	1.000	0.921	0.924	0.922
	4 Character Gram	465	80	116	19.8	0	0.2	0.854	1.000	0.921	0.924	0.921
	3 Character Gram	465	80	116	20	0	0	0.853	1.000	0.921	0.921	0.921
	Unigram	465	80	116	19.9	0	0.1	0.854	1.000	0.921	0.924	0.921
Teens vs. Adults	Trigram	465	1263	0.5	4.7	115.5	311.3	0.077	0.004	0.000	0.017	0.008
	Bigram	465	1263	0.2	1.1	115.8	314.9	0.150	0.002	0.000	0.017	0.002
	3 Character Gram	465	1263	0	0	116	316	0.000	0.000	0.000	0.000	0.000
	4 Character Gram	465	1263	0	0	116	316	0.000	0.000	0.000	0.000	0.000
	5 Character Gram	465	1263	0	0.3	116	315.7	0.000	0.000	0.000	0.000	0.000
	Unigram	465	1263	0	0	116	316	0.000	0.000	0.000	0.000	0.000

Table 69. Naïve Bayes Classifier Results Ranked by Average F-score for Each Classification Task (Laplace Smoothing with Punctuation).

Classification Task	Feature	Teen Training Files	Adult Training Files	True Positives	False Positives	False Negatives	True Negatives	Precision	Recall	Low F-Score	High F-Score	Average F-Score
Teens vs. 20s	Trigram	465	689	3.4	12.2	112.6	159.8	0.182	0.029	0.000	0.127	0.047
	Bigram	465	689	0.4	3.8	115.6	168.2	0.079	0.003	0.000	0.033	0.004
	3 Character Gram	465	689	0	0	116	172	0.000	0.000	0.000	0.000	0.000
	4 Character Gram	465	689	0	0.2	116	171.8	0.000	0.000	0.000	0.000	0.000
	5 Character Gram	465	689	0	0.9	116	171.1	0.000	0.000	0.000	0.000	0.000
	Unigram	465	689	0	0.1	116	171.9	0.000	0.000	0.000	0.000	0.000
Teens vs. 30s	Trigram	465	259	105.2	27.1	10.8	37.9	0.786	0.907	0.440	0.910	0.875
	Bigram	465	259	82.2	24.9	33.8	40.1	0.730	0.709	0.146	0.891	0.744
	Unigram	465	259	26.7	11.9	89.3	53.1	0.475	0.230	0.000	0.901	0.235
	5 Character Gram	465	259	23.7	7.7	92.3	57.3	0.507	0.204	0.000	0.935	0.205
	4 Character Gram	465	259	16.7	5.5	99.3	59.5	0.385	0.144	0.000	0.922	0.097
	3 Character Gram	465	259	7.8	2.1	108.2	62.9	0.167	0.067	0.000	0.702	0.023
Teens vs. 40s	5 Character Gram	465	235	104.6	7.5	11.4	51.5	0.926	0.902	0.405	0.979	0.949
	Unigram	465	235	105.1	10.5	10.9	48.5	0.900	0.906	0.338	0.971	0.943
	4 Character Gram	465	235	98.9	6.8	17.1	52.2	0.905	0.853	0.081	0.983	0.926
	Bigram	465	235	115.6	21.4	0.4	37.6	0.844	0.997	0.898	0.928	0.914
	Trigram	465	235	115.2	25.9	0.8	33.1	0.817	0.993	0.880	0.913	0.896
	3 Character Gram	465	235	88	6.6	28	52.4	0.873	0.759	0.017	0.983	0.814
Teens vs. 50s	5 Character Gram	465	80	116	19	0	1	0.859	1.000	0.921	0.932	0.924
	Bigram	465	80	116	19.7	0	0.3	0.855	1.000	0.921	0.924	0.922
	4 Character Gram	465	80	116	19.6	0	0.4	0.855	1.000	0.921	0.928	0.922
	Trigram	465	80	115.4	19.1	0.6	0.9	0.858	0.995	0.916	0.928	0.921
	3 Character Gram	465	80	116	19.9	0	0.1	0.854	1.000	0.921	0.924	0.921
	Unigram	465	80	116	19.9	0	0.1	0.854	1.000	0.921	0.924	0.921
Teens vs. Adults	Trigram	465	1263	0.6	4	115.4	312	0.130	0.005	0.000	0.049	0.006
	Bigram	465	1263	0.2	1.1	115.8	314.9	0.150	0.002	0.000	0.017	0.002
	3 Character Gram	465	1263	0	0	116	316	0.000	0.000	0.000	0.000	0.000
	4 Character Gram	465	1263	0	0	116	316	0.000	0.000	0.000	0.000	0.000
	5 Character Gram	465	1263	0	0.6	116	315.4	0.000	0.000	0.000	0.000	0.000
	Unigram	465	1263	0	0	116	316	0.000	0.000	0.000	0.000	0.000

Table 70. Naïve Bayes Classifier Results Ranked by Average F-score for Each Classification Task (Laplace Smoothing without Punctuation).

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E: SUPPORT VECTOR MACHINE RESULTS

This appendix contains the Support Vector Machine results. The results for each classification task are ranked by average F-score. In order to exclude outliers, the average F-score was calculated without the highest and lowest F-score measure from the 10 random test sets. The lowest and highest F-score results, however, are included in the tables as separate columns. The average number of true positives, false positives, true negatives, false negatives, precision and recall do not omit the highest/lowest respective value from the 10 random test set results. Tables 71–75 display the results for each classification task.

Feature	Slack Variable	True Positives	True Negatives	False Positives	False Negatives	Precision	Recall	Low F-score	High F-score	Average F-score
Trigram (25 entropy)	0.000030500	90.1	149.4	37.0	25.9	0.694	0.777	0.000	0.987	0.769
Trigram (15 entropy)	0.000030500	89.7	149.6	36.8	26.3	0.695	0.773	0.000	0.987	0.765
Trigram (5 entropy)	0.000061000	87.3	150.5	35.9	28.7	0.726	0.753	0.000	0.987	0.760
Trigram (50 entropy)	0.000030500	90.9	138.6	47.8	25.1	0.685	0.784	0.000	0.987	0.754
Trigram	0.000030500	90.4	137.1	49.3	25.6	0.675	0.779	0.000	0.987	0.744
Trigram (75 entropy)	0.000030500	90.9	128.4	58.0	25.1	0.676	0.784	0.000	0.987	0.741
Bigram (5 mutual)	0.000030500	76.2	168.2	18.2	39.8	0.726	0.657	0.000	0.991	0.694
Unigram (50 mutual)	0.000030500	76.9	152.8	33.6	39.1	0.705	0.663	0.000	0.987	0.663
Bigram (50 mutual)	0.000030500	72.6	163.2	23.2	43.4	0.705	0.626	0.000	0.966	0.656
Bigram (15 mutual)	0.000030500	73.2	156.2	30.2	42.8	0.674	0.631	0.000	0.862	0.638
Bigram (75 mutual)	0.000030500	75.8	143.0	43.4	40.2	0.670	0.653	0.000	0.938	0.628
Bigram (75 entropy)	0.000030500	75.2	148.2	38.2	40.8	0.642	0.648	0.000	0.987	0.625
Unigram (25 mutual)	0.000030500	70.4	164.8	21.6	45.6	0.662	0.607	0.000	0.963	0.615
Bigram (50 entropy)	0.000030500	77.1	141.2	45.2	38.9	0.619	0.665	0.000	0.987	0.614
Bigram	0.000030500	78.2	137.8	48.6	37.8	0.614	0.674	0.000	0.987	0.611
Trigram (75 mutual)	0.000030500	76.0	123.9	62.5	40.0	0.691	0.655	0.184	0.897	0.609
Bigram (25 entropy)	0.000030500	78.0	130.7	55.7	38.0	0.603	0.672	0.000	0.987	0.597
Bigram (25 mutual)	0.000061000	68.1	166.1	20.3	47.9	0.663	0.587	0.000	0.983	0.591
Bigram (5 entropy)	0.000030500	78.2	119.9	66.5	37.8	0.594	0.674	0.000	0.987	0.585
Bigram (15 entropy)	0.000030500	78.1	119.4	67.0	37.9	0.594	0.673	0.000	0.987	0.584
Trigram (5 mutual)	0.000030500	64.1	153.2	33.2	51.9	0.681	0.553	0.000	0.970	0.581
Trigram (15 mutual)	0.000030500	68.6	146.6	39.8	47.4	0.586	0.591	0.017	0.970	0.576
Unigram (75 mutual)	0.000122070	61.0	139.6	46.8	55.0	0.775	0.526	0.083	0.825	0.551
Unigram (75 entropy)	0.000030500	70.5	122.5	63.9	45.5	0.576	0.608	0.099	0.762	0.545
Trigram (meta-data)	0.000030500	69.3	132.0	54.4	46.7	0.669	0.597	0.033	0.954	0.544
3 Character Gram (meta-data)	0.000244141	71.8	109.9	76.5	44.2	0.584	0.619	0.014	0.971	0.540
Trigram (50 mutual)	0.000030500	74.6	102.4	84.0	41.4	0.565	0.643	0.065	0.952	0.537
Unigram	0.000030500	73.7	113.1	73.3	42.3	0.551	0.635	0.099	0.762	0.535
Unigram (5 entropy)	0.000030500	73.6	113.1	73.3	42.4	0.551	0.634	0.099	0.762	0.534
Trigram (25 mutual)	0.000030500	61.8	150.2	36.2	54.2	0.618	0.533	0.111	0.940	0.524
Unigram (25 entropy)	0.000030500	73.5	106.0	80.4	42.5	0.540	0.634	0.099	0.762	0.521
Bigram (meta-data)	0.000061000	64.5	140.1	46.3	51.5	0.602	0.556	0.000	0.946	0.516
Unigram (15 mutual)	0.000030500	67.7	131.4	55.0	48.3	0.625	0.584	0.017	0.963	0.515
Unigram (15 entropy)	0.000030500	73.5	96.6	89.8	42.5	0.532	0.634	0.099	0.762	0.509
Unigram (50 entropy)	0.000030500	73.5	93.8	92.6	42.5	0.529	0.634	0.099	0.762	0.505
Unigram (5 mutual)	0.000061000	61.0	148.7	37.7	55.0	0.682	0.526	0.079	0.978	0.502
3 Character Gram	0.000122070	72.6	90.6	95.8	43.4	0.525	0.626	0.021	0.872	0.494
Unigram (meta-data)	0.000122070	62.5	132.9	53.5	53.5	0.573	0.539	0.058	0.987	0.483
Lin Features	128.000000000	54.4	92.4	94.0	61.6	0.420	0.469	0.058	0.558	0.468

Table 71. Teens Versus 20s Support Vector Machine Results (Ranked by Average F-score).

Feature	Slack Variable	True Positives	True Negatives	False Positives	False Negatives	Precision	Recall	Low F-score	High F-score	Average F-score
Bigram (15 mutual)	0.000976563	109.6	46.2	18.8	6.4	0.880	0.945	0.734	0.991	0.914
Bigram (50 mutual)	0.000244141	111.4	45.0	20.0	4.6	0.878	0.960	0.780	1.000	0.913
Unigram (50 mutual)	0.000488281	104.9	43.2	21.8	11.1	0.865	0.904	0.307	1.000	0.894
Bigram (25 mutual)	0.000488281	111.0	41.2	23.8	5.0	0.857	0.957	0.782	1.000	0.893
Bigram (5 mutual)	0.000488281	107.2	45.7	19.3	8.8	0.876	0.924	0.742	0.996	0.891
Trigram (5 mutual)	0.000488281	103.9	46.2	18.8	12.1	0.880	0.896	0.550	0.987	0.888
Unigram (15 mutual)	0.000488281	99.5	47.3	17.7	16.5	0.884	0.858	0.098	1.000	0.888
Bigram (75 mutual)	0.000061000	109.5	42.5	22.5	6.5	0.862	0.944	0.780	1.000	0.888
Trigram (15 mutual)	0.000244141	105.4	44.8	20.2	10.6	0.870	0.909	0.594	0.982	0.888
Unigram (25 mutual)	0.000488281	96.5	50.3	14.7	19.5	0.894	0.832	0.067	0.978	0.886
Trigram (25 mutual)	0.000122070	112.3	36.6	28.4	3.7	0.829	0.968	0.781	0.975	0.885
Bigram (50 entropy)	0.000122070	97.5	44.7	20.3	18.5	0.819	0.841	0.094	1.000	0.884
Unigram (5 mutual)	0.000488281	99.1	43.9	21.1	16.9	0.807	0.854	0.088	1.000	0.883
Trigram (25 entropy)	0.000030500	101.1	42.1	22.9	14.9	0.761	0.872	0.000	0.991	0.882
Bigram (25 entropy)	0.000122070	97.2	44.7	20.3	18.8	0.819	0.838	0.094	1.000	0.881
Trigram (15 entropy)	0.000030500	101.0	42.1	22.9	15.0	0.761	0.871	0.000	0.991	0.881
Unigram (75 mutual)	0.000488281	104.1	38.4	26.6	11.9	0.841	0.897	0.098	1.000	0.878
Unigram (25 entropy)	0.000488281	99.8	44.1	20.9	16.2	0.845	0.860	0.462	1.000	0.865
Trigram (50 entropy)	0.000030500	102.0	36.7	28.3	14.0	0.731	0.879	0.000	0.991	0.863
Trigram (5 entropy)	0.000030500	101.2	37.7	27.3	14.8	0.733	0.872	0.000	0.991	0.863
Trigram	0.000030500	102.0	35.8	29.2	14.0	0.727	0.879	0.000	0.991	0.860
Trigram (75 entropy)	0.000030500	102.0	35.8	29.2	14.0	0.727	0.879	0.000	0.991	0.860
Bigram (75 entropy)	0.000122070	97.5	39.0	26.0	18.5	0.787	0.841	0.094	1.000	0.860
Unigram (75 entropy)	0.000488281	100.0	42.0	23.0	16.0	0.837	0.862	0.462	1.000	0.858
Unigram (50 entropy)	0.000488281	100.0	42.0	23.0	16.0	0.836	0.862	0.462	1.000	0.858
Bigram (15 entropy)	0.000122070	97.5	38.4	26.6	18.5	0.785	0.841	0.094	1.000	0.858
Bigram (5 entropy)	0.000122070	97.5	38.4	26.6	18.5	0.785	0.841	0.094	1.000	0.858
Unigram (15 entropy)	0.000488281	99.9	41.5	23.5	16.1	0.835	0.861	0.462	1.000	0.856
Unigram	0.000488281	99.6	41.6	23.4	16.4	0.834	0.859	0.462	1.000	0.855
Unigram (5 entropy)	0.000488281	99.6	41.6	23.4	16.4	0.834	0.859	0.462	1.000	0.855
Bigram	0.000122070	96.0	40.0	25.0	20.0	0.790	0.828	0.094	1.000	0.852
Trigram (75 mutual)	0.015625000	94.9	45.0	20.0	21.1	0.764	0.818	0.000	0.996	0.852
Trigram (50 mutual)	0.000030500	100.9	37.3	27.7	15.1	0.810	0.870	0.254	1.000	0.850
3 Character Gram	0.000061000	95.3	40.2	24.8	20.7	0.740	0.822	0.015	1.000	0.839
3 Character Gram (meta-data)	0.000061000	95.4	34.4	30.6	20.6	0.710	0.822	0.015	1.000	0.816
Bigram (meta-data)	0.000122070	91.1	36.5	28.5	24.9	0.759	0.785	0.097	1.000	0.808
Unigram (meta-data)	0.000061000	95.4	30.9	34.1	20.6	0.750	0.822	0.295	0.987	0.794
Lin Features	0.000061000	115.0	2.8	62.2	1.0	0.649	0.991	0.777	0.789	0.785
Trigram (meta-data)	0.000030500	90.5	29.7	35.3	25.5	0.750	0.780	0.050	0.946	0.765

Table 72. Teens Versus 30s Support Vector Machine Results (Ranked by Average F-score).

Feature	Slack Variable	True Positives	True Negatives	False Positives	False Negatives	Precision	Recall	Low F-score	High F-score	Average F-score
Trigram (75 mutual)	0.000030500	115.4	52.4	6.6	0.6	0.959	0.995	0.808	1.000	0.991
Bigram (75 mutual)	0.000030500	104.6	58.6	0.4	11.4	0.994	0.902	0.461	1.000	0.980
Trigram (25 mutual)	0.000030500	112.4	54.2	4.8	3.6	0.964	0.969	0.836	1.000	0.977
Bigram (50 mutual)	0.000030500	103.6	58.8	0.2	12.4	0.996	0.893	0.430	1.000	0.976
Unigram (75 mutual)	0.000030500	108.6	56.5	2.5	7.4	0.979	0.936	0.710	1.000	0.975
Trigram (15 mutual)	0.000030500	110.3	55.1	3.9	5.7	0.968	0.951	0.792	1.000	0.974
Trigram (50 mutual)	0.000030500	114.1	51.4	7.6	1.9	0.947	0.984	0.836	1.000	0.974
Unigram	0.000030500	103.6	57.7	1.3	12.4	0.987	0.893	0.538	1.000	0.962
Unigram (5 entropy)	0.000030500	103.6	57.7	1.3	12.4	0.987	0.893	0.538	1.000	0.962
Unigram (50 mutual)	0.000030500	101.9	58.1	0.9	14.1	0.989	0.878	0.487	1.000	0.960
Unigram (15 entropy)	0.000030500	103.5	57.4	1.6	12.5	0.985	0.892	0.538	1.000	0.959
Unigram (50 entropy)	0.000030500	103.5	57.3	1.7	12.5	0.984	0.892	0.538	1.000	0.959
Unigram (75 entropy)	0.000030500	103.5	57.3	1.7	12.5	0.984	0.892	0.538	1.000	0.959
Bigram (25 mutual)	0.000061000	108.1	53.5	5.5	7.9	0.965	0.932	0.735	1.000	0.959
Trigram	0.000030500	111.1	52.1	6.9	4.9	0.950	0.958	0.853	1.000	0.957
Trigram (15 entropy)	0.000030500	111.1	52.1	6.9	4.9	0.950	0.958	0.853	1.000	0.957
Trigram (25 entropy)	0.000030500	111.1	52.1	6.9	4.9	0.950	0.958	0.853	1.000	0.957
Unigram (25 entropy)	0.000030500	102.8	57.7	1.3	13.2	0.987	0.886	0.538	1.000	0.957
Trigram (5 entropy)	0.000030500	111.1	52.1	6.9	4.9	0.950	0.958	0.856	1.000	0.957
Trigram (5 mutual)	0.000030500	107.6	56.3	2.7	8.4	0.978	0.928	0.833	1.000	0.956
Trigram (75 entropy)	0.000030500	111.0	51.9	7.1	5.0	0.949	0.957	0.853	1.000	0.956
Trigram (50 entropy)	0.000030500	110.3	52.1	6.9	5.7	0.950	0.951	0.853	1.000	0.953
Bigram (25 entropy)	0.000122070	105.0	54.4	4.6	11.0	0.966	0.905	0.594	1.000	0.952
Bigram	0.000030500	100.1	55.1	3.9	15.9	0.932	0.863	0.094	1.000	0.952
Bigram (50 entropy)	0.000030500	100.1	55.1	3.9	15.9	0.932	0.863	0.094	1.000	0.952
Bigram (15 entropy)	0.000122070	104.7	54.3	4.7	11.3	0.965	0.903	0.568	1.000	0.952
Bigram (5 entropy)	0.000122070	104.7	54.1	4.9	11.3	0.963	0.903	0.568	1.000	0.951
Bigram (75 entropy)	0.000030500	100.1	54.7	4.3	15.9	0.929	0.863	0.094	1.000	0.950
Bigram (15 mutual)	0.000488281	100.9	57.9	1.1	15.1	0.989	0.870	0.541	1.000	0.947
Unigram (25 mutual)	0.000244141	92.6	57.4	1.6	23.4	0.967	0.798	0.127	1.000	0.910
Unigram (15 mutual)	0.000244141	92.6	58.2	0.8	23.4	0.993	0.798	0.173	1.000	0.910
3 Character Gram	0.000244141	96.5	52.6	6.4	19.5	0.931	0.832	0.400	1.000	0.907
3 Character Gram (meta-data)	0.000244141	96.5	52.6	6.4	19.5	0.931	0.832	0.400	1.000	0.907
Bigram (5 mutual)	0.000244141	97.1	56.8	2.2	18.9	0.978	0.837	0.667	1.000	0.902
Bigram (meta-data)	0.000030500	95.1	52.1	6.9	20.9	0.931	0.820	0.378	0.996	0.896
Unigram (meta-data)	0.000030500	93.1	53.5	5.5	22.9	0.942	0.803	0.439	1.000	0.877
Unigram (5 mutual)	0.000030500	89.6	54.9	4.1	26.4	0.970	0.772	0.188	1.000	0.877
Trigram (meta-data)	0.000030500	91.9	47.7	11.3	24.1	0.902	0.792	0.430	1.000	0.846
Lin Features	0.125000000	111.3	17.3	41.7	4.7	0.728	0.959	0.796	0.867	0.827

Table 73. Teens Versus 40s Support Vector Machine Results (Ranked by Average F-score).

Feature	Slack Variable	True Positives	True Negatives	False Positives	False Negatives	Precision	Recall	Low F-score	High F-score	Average F-score
Trigram (50 mutual)	0.000030500	112.0	15.1	20.1	4.0	0.910	0.966	0.589	0.996	0.958
Trigram (25 mutual)	0.000030500	111.3	14.6	20.6	4.7	0.909	0.959	0.583	0.996	0.953
Trigram (75 mutual)	0.000030500	102.6	29.2	6.0	13.4	0.931	0.884	0.366	1.000	0.951
Trigram (15 mutual)	0.000030500	110.3	24.3	10.9	5.7	0.927	0.951	0.736	0.996	0.951
Bigram (25 entropy)	0.001953125	103.5	27.8	7.4	12.5	0.923	0.892	0.388	0.996	0.946
Unigram (75 mutual)	0.000061000	106.8	28.0	7.2	9.2	0.943	0.921	0.643	1.000	0.945
Bigram (75 mutual)	0.000030500	104.9	29.8	5.4	11.1	0.955	0.904	0.594	1.000	0.945
Trigram (5 mutual)	0.000030500	108.1	15.4	19.8	7.9	0.914	0.932	0.583	0.996	0.935
3 Character Gram (meta-data)	0.000976563	105.9	26.7	8.5	10.1	0.934	0.913	0.635	1.000	0.932
Trigram	0.000030500	96.4	32.6	2.6	19.6	0.975	0.831	0.202	1.000	0.925
Lin Features	0.000030500	115.8	1.8	33.4	0.2	0.814	0.998	0.589	0.928	0.922
Bigram (50 entropy)	0.000122070	96.5	30.6	4.6	19.5	0.942	0.832	0.245	0.996	0.921
Bigram (15 entropy)	0.001953125	102.8	21.7	13.5	13.2	0.889	0.886	0.388	0.996	0.917
Unigram (50 entropy)	0.000061000	97.6	31.4	3.8	18.4	0.951	0.841	0.333	1.000	0.914
Bigram (50 mutual)	0.000030500	103.6	23.7	11.5	12.4	0.922	0.893	0.627	0.996	0.913
Bigram (15 mutual)	0.000122070	98.5	29.3	5.9	17.5	0.946	0.849	0.487	1.000	0.913
Bigram	0.001953125	103.2	18.3	16.9	12.8	0.880	0.890	0.388	0.996	0.910
Trigram (25 entropy)	0.000030500	96.7	26.1	9.1	19.3	0.938	0.834	0.202	1.000	0.903
Unigram	0.000976563	103.3	19.2	16.0	12.7	0.908	0.891	0.657	0.991	0.900
Unigram (5 entropy)	0.000976563	103.2	19.1	16.1	12.8	0.908	0.890	0.655	0.991	0.899
Unigram (15 entropy)	0.000976563	104.0	13.4	21.8	12.0	0.898	0.897	0.590	0.991	0.898
Unigram (25 entropy)	0.000976563	98.7	27.6	7.6	17.3	0.933	0.851	0.636	0.991	0.898
Bigram (5 entropy)	0.000030500	103.4	15.5	19.7	12.6	0.912	0.891	0.586	1.000	0.896
Trigram (15 entropy)	0.000030500	95.9	25.9	9.3	20.1	0.936	0.827	0.202	1.000	0.896
Bigram (75 entropy)	0.001953125	102.9	11.8	23.4	13.1	0.868	0.887	0.388	0.991	0.895
Unigram (50 mutual)	0.000061000	103.3	14.3	20.9	12.7	0.904	0.891	0.522	0.996	0.894
3 Character Gram	0.000488281	98.0	27.2	8.0	18.0	0.930	0.845	0.519	0.967	0.892
Trigram (5 entropy)	0.000030500	95.8	16.6	18.6	20.2	0.914	0.826	0.202	1.000	0.872
Bigram (25 mutual)	0.000030500	90.3	29.8	5.4	25.7	0.919	0.778	0.050	0.996	0.871
Trigram (50 entropy)	0.000030500	95.5	16.6	18.6	20.5	0.915	0.823	0.202	1.000	0.870
Trigram (75 entropy)	0.000030500	94.4	19.0	16.2	21.6	0.921	0.814	0.202	1.000	0.865
Unigram (25 mutual)	0.000030500	91.3	25.3	9.9	24.7	0.916	0.787	0.159	0.996	0.848
Unigram (5 mutual)	0.000030500	91.9	16.7	18.5	24.1	0.911	0.792	0.188	0.991	0.847
Unigram (meta-data)	0.000030500	86.6	26.6	8.6	29.4	0.882	0.747	0.092	0.996	0.823
Unigram (75 entropy)	0.000030500	92.4	13.9	21.3	23.6	0.886	0.797	0.188	0.996	0.823
Unigram (15 mutual)	0.000030500	90.4	15.7	19.5	25.6	0.872	0.779	0.162	0.996	0.805
Bigram (meta-data)	0.000030500	89.9	13.6	21.6	26.1	0.887	0.775	0.435	1.000	0.800
Trigram (meta-data)	0.000030500	82.2	15.3	19.9	33.8	0.899	0.709	0.307	0.983	0.760
Bigram (5 mutual)	0.000030500	84.7	15.1	20.1	31.3	0.867	0.730	0.120	1.000	0.738

Table 74. Teens Versus 50s Support Vector Machine Results (Ranked by Average F-score).

Feature	Slack Variable	True Positives	True Negatives	False Positives	False Negatives	Precision	Recall	Low F-score	High F-score	Average F-score
Unigram (5 mutual)	0.000030500	89.5	253.7	32.7	26.5	0.846	0.772	0.114	0.991	0.786
Bigram (15 mutual)	0.000030500	95.2	231.9	54.5	20.8	0.709	0.821	0.033	0.996	0.766
Bigram (5 mutual)	0.003906250	95.2	195.8	90.6	20.8	0.686	0.821	0.038	1.000	0.757
Trigram	0.000061000	100.9	196.3	90.1	15.1	0.647	0.870	0.025	1.000	0.756
Trigram (5 entropy)	0.000061000	100.9	195.0	91.4	15.1	0.644	0.870	0.025	1.000	0.754
Trigram (75 entropy)	0.000061000	100.7	196.3	90.1	15.3	0.646	0.868	0.026	1.000	0.753
Trigram (50 entropy)	0.000061000	99.3	196.0	90.4	16.7	0.645	0.856	0.026	1.000	0.745
Trigram (5 mutual)	0.000030500	99.2	201.6	84.8	16.8	0.682	0.855	0.167	0.991	0.743
Trigram (15 entropy)	0.000061000	97.9	195.8	90.6	18.1	0.646	0.844	0.025	1.000	0.737
Unigram (15 mutual)	0.000030500	87.5	249.3	37.1	28.5	0.742	0.754	0.000	1.000	0.735
Trigram (25 entropy)	0.000122070	95.3	181.7	104.7	20.7	0.654	0.822	0.000	1.000	0.734
Trigram (50 mutual)	0.000030500	96.5	204.6	81.8	19.5	0.778	0.832	0.294	0.991	0.732
Bigram (25 mutual)	0.000030500	98.6	201.6	84.8	17.4	0.648	0.850	0.031	1.000	0.729
Bigram (50 mutual)	0.000244141	94.2	169.6	116.8	21.8	0.671	0.812	0.014	1.000	0.728
Trigram (15 mutual)	0.000030500	93.1	206.5	79.9	22.9	0.686	0.803	0.044	1.000	0.721
Unigram (75 mutual)	0.000030500	90.5	224.8	61.6	25.5	0.689	0.780	0.000	1.000	0.720
Bigram (75 mutual)	0.000030500	99.1	187.3	99.1	16.9	0.631	0.854	0.000	1.000	0.716
Trigram (25 mutual)	0.000030500	96.2	201.8	84.6	19.8	0.749	0.829	0.294	0.987	0.708
Trigram (75 mutual)	0.000030500	92.5	203.8	82.6	23.5	0.782	0.797	0.294	1.000	0.705
Bigram (25 entropy)	0.000030500	96.0	196.3	90.1	20.0	0.727	0.828	0.294	1.000	0.700
Unigram (25 mutual)	0.000030500	85.0	238.8	47.6	31.0	0.702	0.733	0.000	1.000	0.698
Bigram (50 entropy)	0.000030500	96.5	193.4	93.0	19.5	0.723	0.832	0.294	1.000	0.698
Bigram (5 entropy)	0.000030500	96.0	191.9	94.5	20.0	0.722	0.828	0.294	1.000	0.694
Bigram	0.000030500	96.7	188.4	98.0	19.3	0.716	0.834	0.294	1.000	0.692
Bigram (75 entropy)	0.000030500	96.1	192.3	94.1	19.9	0.711	0.828	0.294	1.000	0.688
Bigram (15 entropy)	0.000030500	96.0	190.3	96.1	20.0	0.710	0.828	0.294	1.000	0.686
Unigram (50 mutual)	0.000030500	85.5	229.7	56.7	30.5	0.683	0.737	0.000	0.996	0.679
Unigram (50 entropy)	0.000030500	85.0	228.9	57.5	31.0	0.713	0.733	0.174	1.000	0.670
Unigram (5 entropy)	0.000030500	84.7	229.1	57.3	31.3	0.707	0.730	0.138	0.996	0.670
Unigram (15 entropy)	0.000030500	84.5	229.0	57.4	31.5	0.702	0.728	0.112	0.996	0.670
Unigram	0.000030500	84.6	229.1	57.3	31.4	0.706	0.729	0.138	0.996	0.669
Unigram (25 entropy)	0.000030500	82.8	228.9	57.5	33.2	0.702	0.714	0.112	0.996	0.660
Unigram (75 entropy)	0.000030500	84.5	227.0	59.4	31.5	0.692	0.728	0.138	0.996	0.659
Unigram (meta-data)	0.000030500	81.3	224.7	61.7	34.7	0.680	0.701	0.067	1.000	0.640
Trigram (meta-data)	0.000244141	91.1	170.0	116.4	24.9	0.540	0.785	0.029	0.996	0.640
Bigram (meta-data)	0.000030500	85.8	189.7	96.7	30.2	0.651	0.740	0.191	0.996	0.584
3 Character Gram (meta-data)	0.001953125	72.7	192.9	93.5	43.3	0.633	0.627	0.067	1.000	0.543
3 Character Gram	0.001953125	72.7	191.5	94.9	43.3	0.622	0.627	0.067	0.960	0.541
Lin Features	256.000000000	41.9	233.0	53.4	74.1	0.432	0.361	0.199	0.921	0.327

Table 75. Teens Versus Adults Support Vector Machine Results (Ranked by Average F-score).

APPENDIX F: LIN FEATURES

This appendix contains the feature dictionary for the Lin Feature Set [5]. Her feature vector contains not only the average sentence length and average number of word types, but it also contains the type count and token count for each punctuation mark/emoticon. Figure 7 contains the Lin feature dictionary.

!	:	:tongue:	"	[	\
#	:-@	;	$	]	~
%	:-(	;-)	(	^	Average Sentence Length
&	:-)	<	)	|	Average Number of Word Types
	:-o	=	*	_	
,	:beer:	>	+	`	
-	:blush:	>:->	.	{	
/	:love:	@	?	}	

Figure 7. Lin Feature Dictionary [After 5].

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX G: EMOTICON DICTIONARY

This appendix contains the emoticon dictionary used. All, but four emoticons are from the Wikipedia website [23]. The four emoticons, which begin and end with a colon, are from the built-in emoticons in the NPS Chat Corpus. Figure 8 displays the emoticon dictionary used in this study.

:)	8)	:O	:-D	>:)	8-)	(=C
:-)	;)	:/	:S	}:)	8-0	>o>
:^)	*)	:-/	:s	0:)	:'-(	<o<
=)	X(	:-\	^o)	<3	:-*	O<3=
B)	(|)	:\	:3	x3	X-(	:beer:
c8	:(	8/	>:3	</3	:-&	:blush:
cB	8c	8\	:E	:'(	;^)	:love:
=]	Bc	>/	:F	:,(	:-}	:tongue:
:]	B(	>\	:X	:_(	<:}	
x]	|8c	:|	:-*	:*(	>:L	
:D	|8C	:I	>:O	:…(	:9	
e.e	:[	xP	XO	@}->--	*\o/*	
o.oU	:P	XP	:-(o)	--^-@	>:D	
38*	:p	xD	>:(	%-)	>=D	
D:	|D	XD	>[	%-(	:0->-<|:	

Figure 8. Emoticon Dictionary.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX H: PUNCTUATION DICTIONARY

This appendix contains the punctuation dictionary used. The punctuation marks are the non-alphanumeric keys on the QWERTY keyboard. Figure 9 displays the punctuation dictionary used in this study.

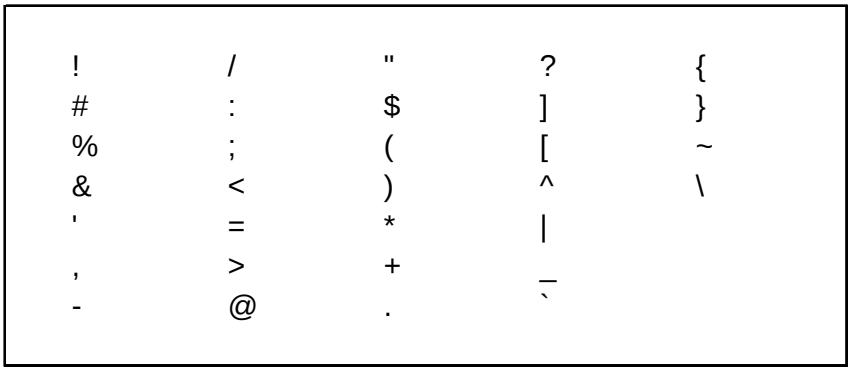

Figure 9. Punctuation Dictionary.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] A. Lenhart, M. Madden, A.R. Macgill and A. Smith, "Teens and Social Media: The use of social media gains a greater foothold in teen life as they embrace the conversational nature of interactive online media," vol. 2009, pp. 36, 2007.

[2] J. Wolak, D. Finkelhor and K.J. Mitchell, *Online Victimization of Youth: Five Years Later.* Alexandria, Va.: National Center for Missing & Exploited Children, pp. 79, 2006.

[3] K. Mitchell, D. Finkelhor and J. Wolak, "Youth Internet Users at Risk for the Most Serious Online Sexual Solicitations," *American Journal of Preventive Medicine,* vol. 32, pp. 532, 2007.

[4] N. Pendar, "Toward Spotting the Pedophile Telling victim from predator in text chats," *International Conference on Semantic Computing, 2007,* pp. 235–241, 2007.

[5] J. Lin, "Automatic Author Profiling of Online Chat Logs," M.S. thesis, Naval Postgraduate School, Monterey, CA, 2007.

[6] H. Dong, S. C. Hui and Y. He, "Structural analysis of chat messages for topic detection," *Online Information Review,* vol. 30, pp. 496–516, 2006.

[7] H. Baayen, H. van Halteren and F. Tweedie, "Outside the cave of shadows: using syntactic annotation to enhance authorship attribution," *Literary and Linguistic Computing,* vol. 11, pp. 121–132, September 1, 1996.

[8] P. Rayson, G. Leech and M. Hodges, "Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus," *International Journal of Corpus Linguistics,* vol. 2, pp. 133–152, 1997.

[9] V. Savicki, D. Lingenfelter and M. Kelley, "Gender Language Style and Group Composition in Internet Discussion Groups," *Journal of Computer-Mediated Communication,* vol. 2, pp. 1, 1996.

[10] E. Alpaydin, *Introduction to Machine Learning.* Cambridge, Mass.: MIT Press, 2004.

[11] T. Kucukyilmaz, B.B. Cambazoglu, C. Aykanat and F. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication," *Information Processing Management,* vol. 44, pp. 1448–1466, 2008.

[12] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information Science and Technology,* vol. 60, pp. 538–556, 2009.

[13] R. Zheng, J. Li, H. Chen and Z. Huang, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," *Journal of the American Society for Information Science and Technology,* vol. 57, pp. 378, 2006.

[14] D. Jurafsky and J.H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* London: Prentice Hall, Pearson Education International, 2009.

[15] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing.* Cambridge, Mass.: MIT Press, 6th edition, 2003.

[16] J. Rudman, "The State of Authorship Attribution Studies: Some Problems and Solutions," *Computers and the Humanities,* vol. 31, pp. 351–365, 1997.

[17] I. H. Witten and T.C. Bell, "The Zero-Frequency Problem—Estimating the Probabilities of Novel Events in Adaptive Text Compression," *IEEE Transactions on Information Theory,* vol. 37, pp. 1085, 1991.

[18] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121–167, 1998.

[19] R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research,* vol. 9, pp. 1871–1874, 2008.

[20] C. Hsu, C. Chang and C. Lin. (2 October 2008), A practical guide to support vector classification. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed September 16, 2009).

[21] J.P. Lewis. (2004), A short SVM (support vector machine) tutorial. [Online]. Available: http://scribblethink.org/Work/Notes/svmtutorial.pdf (accessed September 16, 2009).

[22] Wikipedia. (14 September 2009). Support vector machine. [Online]. Available: http://en.wikipedia.org/wiki/Support_vector_machine (accessed September 16, 2009).

[23] Wikipedia. (8 July 2009). List of emoticons. [Online]. Available: http://en.wikipedia.org/wiki/List_of_emoticons (accessed 8 July 2009).

[24] T. M. Mitchell, *Machine Learning.* New York: McGraw-Hill, 1997.

[25] J. Wolak, D. Finkelhor and K. Mitchell, *Child-Pornography Possessors Arrested in Internet-Related Crimes: Findings from the National Juvenile Online Victimization Study.* Alexandria, VA: National Center for Missing & Exploited Children, pp. 47, 2005.

[26] M. Wells, D. Finkelhor, J. Wolak and K. Mitchell, "Law Enforcement Challenges in Internet Child Pornography Crimes," *Sex Offender Law Report,* vol. 5, pp. 41–42, 49, 2004.

[27] R. D. De Veaux, P. F. Velleman and D. E. Bock, *Stats: Data and Models.* Boston: Addison/Wesley, 1st Edition, 2005.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Dr. Craig Martell
   Naval Postgraduate School
   Monterey, California

4. Jenny Tam
   Naval Postgraduate School
   Monterey, California